TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

ARVUTITEADUSE INSTITUUT

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

Samaaegselt kromatiini avatust ja splaissimist mõjutavate geneetiliste variantide leidmine

Bakalaureusetöö

12 EAP

Evelin Aasna

Juhendaja PhD Kaur Alasoo

TARTU 2019

Samaaegselt kromatiini avatust ja splaissimist mõjutavate geneetiliste variantide leidmine

Kokkuvõte. Enamuses inimese geenides leiab aset splaissimine, millega suurendatakse võimalike valkude mitmekesisust. Splaissimise geneetiline kontroll toimub läbi keerulise RNA ja regulatoorsete valkude võrgustiku. Eelnevates analüüsides on kindlaks tehtud avatud kromatiini ja splaissimise seos, mis mõnel juhul on vahendatud faktorite CTCF ja HP1 seondumisest kromatiinile. Uurisin kas CTCF või HP1 seondumine on põhiline mehhanism, mille kaudu kromatiini avatus määrab splaissimise. Analüüsis kasutasin ATAC, RNA ja ChIP sekveneerimise andmeid. Leidsin geneetilisi variante, mis mõjutavad nii kromatiini avatust, splaissimist kui ka CTCF seondumist. Siiski ei tundu CTCF ega HP1 seondumine olevat põhiline mehhanism kromatiini avatuse vahendatud splaissimise kontrolliks. Splaissimisega seotud regulatoorsete alade jätkuv uurimine on oluline, sest splaissimisvigu on seostatud paljude geneetiliste haigustega.

Märksõnad: ATAC-seq, avatud kromatiin, splaissimine, CTCF, HP1.

CERCS teaduseriala: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika, B220 Geneetika, tsütogeneetika.

Discovering genetic variants that affect both chromatin accessibility and RNA splicing

Abstract. Splicing occurs in the majority of human genes. It is a process that increases the variability of possible protein products. Splicing is genetically regulated by an intricate network of proteins and RNA molecules. Previous studies have shown that accessible chromatin and RNA splicing might be linked through binding of CTCF and HP1. I used ATAC, RNA and ChIP sequencing data to find out if binding of CTCF or HP1 to accessible chromatin has an important role in splicing regulation. I found genetic variants that affect chromatin accessibility, splicing and CTCF binding. However, CTCF and HP1 do not seem to have a central role in mediating splicing regulation through binding to DNA. Continued research on splicing associated regulatory regions is important because mutations affecting splicing have been implicated in many genetic diseases.

Keywords: ATAC-seq, accessible chromatin, splicing, CTCF, HP1.

CERCS research specialization: B110 Bioinformatics, medical informatics, biomathematics, biometrics, B220 Genetics, cytogenetics.

Sisukord

Sisukord **3**

Kasutatud lühendid **4**

Sissejuhatus **5**

1. Kirjanduse ülevaade **6**

1.1. Lookuste mõju 6

1.1.1. Kvantitatiivse tunnuse lookus 6

1.1.2. Lookuste distaalne ja lokaalne mõju 6

1.1.3. Lookuste cis ja trans mõju 7

1.1.4. Juhtvariant ja põhjuslik variant 7

1.2. Kromatiini avatus 8

1.2.1. Regulatoorsed alad ja avatud kromatiin 8

1.2.2. Kromatiini avatuse ja geenikspressiooni seos 9

1.3. Splaissimine 10

1.3.1. Geenikspressiooni ja splaissimist mõjutavad erinevad geneetilised variandid 10

1.3.2. Alternatiivne splaissimine ja selle seos haigustega 10

1.3.3. Kromatiini avatuse ja splaissimise seos 11

1.3.4. Splaissimise regulatsioonimehhanismid 11

1.4. Mõõtmistehnoloogiad 12

1.4.1. Kromatiini avatus 12

1.4.2. Transkriptsioonifaktorite seondumine 14

1.4.3. Transkriptsioon 15

2. Eksperimentaalosa **15**

2.1. Töö eesmärgid 15

2.2. Materjal 16

2.2.1. Kumasaka et al. 16

2.2.2. GEUVADIS 16

2.2.3. HP1 ja CTCF ChIP-seq 17

2.2.4. CTCF QTLid 17

2.3. Meetodid 18

2.3.1. DNA ligipääsetavuse analüüs 18

2.3.2. Geenikspressiooni ja splaissimise analüüs 20

2.3.3. Geneetiliste seoste leidmine 21

2.3.4. Kromatiini ligipääsetavuse QTLid 22

2.3.5. RNA QTL 24

2.3.6. CTCF QTL 24

2.3.7. Geneetiliste seoste ülekate 24

2.4. Tulemused 26
2.4.1. Seosed kromatiini ligipääsetavuse ja transkriptsiooni vahel 26
2.4.2. CTCFi ja HP1 mõju splaissimisele 31
Kokkuvõte **37**
Summary **38**
Kirjanduse loetelu **40**
Kasutatud veebiaadressid **44**
Lisad **44**
Lihtlitsents **46**

Kasutatud lühendid

ATAC-seq - transposaas ligipääsetava kromatiini analüüs ja sekveneerimine (ingl *assay for transposase-accessible chromatin using sequencing)*

bp - aluspaar (ingl *basepair*)

caQTL - kromatiini ligipääsetavuse kvantitatiivse tunnuse lookus (ingl *chromatin accessibility QTL*)

CTCF - CCCTC-seonduv faktor (ingl *CCCTC binding factor*)

ChIP-seq - valk-DNA interaktsioonide sekveneerimine (ingl *chromatin immunoprecipitation sequencing*)

CQN - kvantiilnormaliseerimine (ingl *conditional quantile normalization*)

DNAse-seq - DNase I ülitundlike piirkondade sekveneerimine (ingl *DNase I hypersensitive sites sequencing*)

FDR - valeavastuste määr (ingl *false discovery rate*)

FPM - fragmenti miljoni kohta (ingl *fragments per million*)

GWAS - ülegenoomne assotsiatsiooniuuring (ingl *genome-wide association study*)

HP1 - heterokromatiin valk 1 (ingl *heterochromatin protein 1*)

kb - kiloaluspaar, 1000 aluspaari (ingl *kilobase*)

LCL - Epstein-Barri viirusega nakatatud B-rakk (ingl *lymphoblastoid cell line*)

LD - lookuste seotud pärandumine ja selle määr (ingl *linkage disequilibrium*)

QTL - kvantitatiivse tunnuse lookus (ingl *quantitative trait locus*)

sQTL - splaissimise kvantitatiivse tunnuse lookus (ingl *splicing QTL*)

VCF - failiformaat, mis salvestab genotüübid ja genoomsed positsioonid (ingl *variant call format*)

Sissejuhatus

Sekveneerimistehnoloogiate arenguga on hüppeliselt kasvanud kättesaadavate sekveneerimisandmete hulk. Genoomi uurimise keskne küsimus on, kuidas erinevused genotüübis väljenduvad organismi fenotüübiliste tunnustena. Geenide avaldumise seaduspärasuste mõistmine on vajalik, et välja selgitada komplekstunnuste, seal hulgas haiguste tekkemehhanisme.

Geenikespressiooni kontroll on vahendatud DNA regulatoorsete alade poolt läbi sinna seonduvate transriptsioonifaktorite. Transkriptsioonifaktorite seondumiskohad jäävad nukleosoomidest vabaks. Kasutades kromatiini avatuse ja RNA sekveneerimisandmeid on võimalik leida geneetilisi variante, mis mõjutavad nii kromatiini avatust kui geenikespressiooni. Need geneetilised variandid võivad mõjutada geenikespressiooni või splaissimist läbi transkriptsioonifaktorite seondumisvõime muutmise. Võrreldes leitud geneetilisi variante teadaolevate transkriptsioonifaktorite seondumissaitidega on võimalik leida tõendeid konkreetsetest ekspressiooni või splaissimise regulatsioonimehhanismidest.

Käesolevas töös on kirjanduse põhjal antud ülevaade kromatiini avatuse, geenikespressiooni ja splaissimise seoste uurimiseks eelnevalt tehtud analüüsidest. Kirjeldatud on transkriptsioonifaktorite CCCTC-seonduva faktori (CTCF) ja heterokromatiini valk 1 (HP1) seni leitud rolli splaissimise regulatsioonil. Samuti kirjeldatakse sekveneerimismeetodeid, mis võimaldavad kindlaks teha kromatiini avatuse, geenikespressiooni ja transkriptsioonifaktorite seondumise.

Käesoleva töö eksperimentaalosas on analüüsitud 91 indiviidi kromatiini avatuse ja 358 indiviidi geenikespressiooni andmeid. Töö eesmärgiks on leida geneetilisi variante, mis mõjutavad nii kromatiini avatust kui ka splaissimist. Lisaks on vaadeldud CTCF ja HP1 seondumist. Töös on testitud, kas faktorite CTCF ja HP1 seondumine on oluline mehhanism, mille kaudu reguleeritakse alternatiivset splaissimist. Töö on teostatud Tartu Ülikooli arvutiteaduse instituudi bioinformaatika ja andmekaeve töörühmas.

1. Kirjanduse ülevaade

1.1. Lookuste mõju

1.1.1. Kvantitatiivse tunnuse lookus

Kvantitatiivse tunnuse lookus (QTL: *quantitative trait locus*) on geneetiline variant, mille varieeruvusel on statistiline seos uuritava tunnusega. Molekulaarse kvantitatiivse tunnusena on võimalik vaadelda näiteks kromatiini avatust, geenieksпressiooni taset ja valkude seondumist kromatiinile. Selliseid molekulaarseid tunnuseid on nimetatud raku fenotüübiks (Kumasaka *et al*., 2016). QTL analüüsile on aluseks populatsioon, milles esineb genotüübiline varieeruvus. Erinevate sekveneerimisandmete põhjal on võimalik mõõta raku fenotüübi tunnuseid ja analüüsida nende seost genotüübiga.

Organismi fenotüübi kujunemine lähtub genotüübist (Johannsen, 1911). Seni on kindlaks tehtud hulgaliselt lookusi, mille varieeruvus korreleerub varieeruvusega fenotüübilistes tunnustes (Martin ja Orgogozo, 2013). Siiski on fenotüübi kujunemine genotüübi põhjal jätkuvalt uurimise all. Komplekstunnused, nagu risk haigestuda diabeeti või südamehaigustesse ei ole põhjustatud ühe geeni pärandumisest (Lander ja Schork, 1994). Selliste haiguste kujunemise mehhanismide mõistmine võib olla aluseks ravi välja töötamisel. Ülegenoomsetes assotsiatsiooniuuringutes (GWAS *genome wide association study*) leitakse korrelatsioon genotüübi ja tunnuse vahel, mitte ühe geeni põhiselt vaid üle kogu genoomi (Welter *et al*., 2014).

1.1.2. Lookuste distaalne ja lokaalne mõju

Olenevalt genoomse lookuse ja mõjutatava tunnuse kaugusest jagatakse QTLd distaalseteks ja lokaalseteks. Vahekaugust võib mõõta aluspaarides (Albert ja Kruglyak, 2015). Täpsed jaotuse piirid olenevad konkreetsest uurimusest. QTLde tuvastamiseks peab läbi viima statistilise testi iga tunnuse ja lookuse paari vahel. Distaalseid lookuseid otsides suureneb läbiviidavate testide arv. Sellega suureneb ka mitmese testimise korrektuur ja väheneb võimekus seoseid tuvastada.

Statistilise võimekuse piirangute tõttu viiakse suures osas inimgenoomi uuringutes läbi ainult tunnuse lähedal asuvate geneetiliste variantide testimine. RNA sekveneerimisandmete analüüsil kasutati 100kb laiust ala iga kvantifitseeritud transkripti tunnuse ümber (Alasoo *et*

al., 2018). Kromatiini avatuse uuringus kasutati 40kb laiust ala avatud kromatiini lõikude ümbruses (Degner *et al*., 2012). Üldiselt on suurema valimi puhul võimalik tuvastada tunnusest kaugemal asuvaid QTLe.

1.1.3. Lookuste cis ja trans mõju

Kvantitatiivse tunnuse lookused võivad tunnuse avaldumist mõjutada läbi erinevate mehhanismide. *Cis* QTLd on lookused, mis mõjutavad tunnust ainult samal kromosoomil, millel need paiknevad. *Cis* QTL mõju saab tuvastada uurides heterosügoote. Sel juhul esineb homoloogilistes kromosoomides erinevus tunnuse avaldumise määras. *Trans* QTLd mõjutavad tunnuse avaldumist mõlemal homoloogilisel kromosoomil. *Trans* QTLd toimivad läbi difuussete faktorite, mis on võimelised mõjutama alleele mõlemal homoloogilisel kromosoomil (Albert ja Kruglyak, 2015). *Trans* mõju puhul esinevad indiviidide vahelised erinevused, kuid mitte indiviidi sisene erinevus alleelide vahel.

Distaalsed lookused mõjutavad geenieksпressiooni enamasti läbi difuussete faktorite mõju (Albert ja Kruglyak, 2015). Geenieksпressiooni QTLde puhul on näidatud, et lookustel võib lähedalasuvatele tunnustele olla nii *cis* kui *trans* mõju (Rockman ja Kruglyak, 2006).

1.1.4. Juhtvariant ja põhjuslik variant

QTL uuringutes testitakse seost tunnuse ja lookuse vahel. Leitakse vähima p-väärtusega lookus ehk juhtvariant. Varieeruvus selles lookuses on kõige tugevamini seotud tunnuse varieeruvusega. Geneetiline variant võib mõjutada kromatiini avatust, geenieksпressiooni ja transleeritud valkude hulka näiteks läbi tranksriptsioonifaktori seondumisvõime muutmise (Albert ja Kruglyak, 2015). Raku fenotüübiliste tunnuste kujunemise mehhanismide mõistmiseks on alustuseks vaja välja selgitada tunnust mõjutav põhjuslik variant.

Põhjusliku variandi leidmine on raskendatud, sest lähestiku paiknevad geneetilised variandid ei pärandu sõltumatult. Põhjuslikku mõju avaldav variant võib olla statistilisel testil leitud juhtvariant või mõni sellega aheldatud variant. Põhjuslike variantide kindlaks tegemisel võib kasutada eeldust, et igal geneetilisel variandil on väike tõenäosus olla põhjuslik variant. Seetõttu on fenotüübi tunnused, mille juhtvariandid on tugevalt aheldatud, suure tõenäosusega mõjutatud ühest põhjuslikust variandist. Geneetiliste variantide aheldatuse määramiseks võib kasutada suurust r^2. Väärtus r^2 näitab kui suure osa ühe lookuse genotüübi varieeruvusest seletab teise lookuse põhjal sobitatud lineaarne mudel.

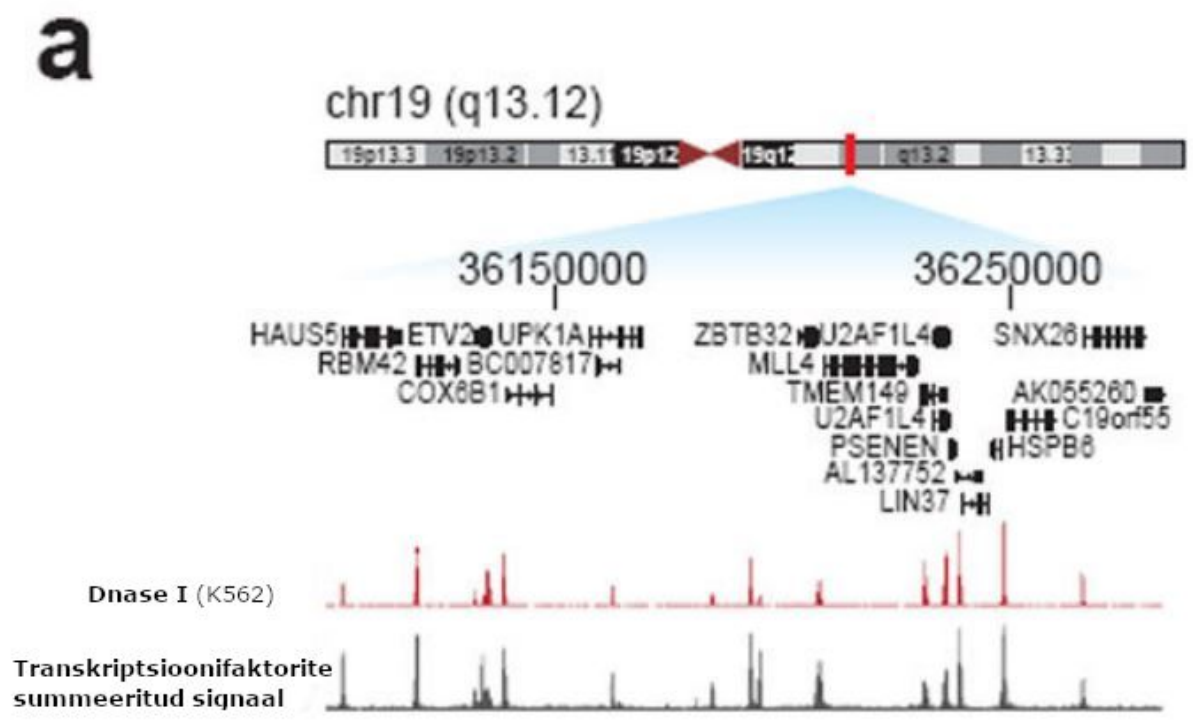
1.2. Kromatiini avatus

1.2.1. Regulatoorsed alad ja avatud kromatiin

Nukleosoomidesse pakkimine on esimene tase kromosoomide efektiivsel ja korrapärasel tuuma paigutamisel (Kornberg, 1974). Kromatiini pakkimisel jäävad avatuks aktiivsed regulatoorsed alad (Gross ja Garrard, 1988). Samas on regulatoorsete alade täpsed biokeemilised funktsioonid jätkuvalt uurimise all. Biokeemiliste funktsioonide väljaselgitamiseks on suur praktiline huvi, kuna enamus GWAS uuringutes haigustega seostatud geneetilised variandid kattuvad regulatoorsete aladega või paiknevad nende lähedal ning ainult väike osa (kuni 10%) asuvad kodeerivates piirkondades (The ENCODE Project Consortium, 2012).

Katseliselt on kindlaks tehtud, et transkriptsioonifaktorite seondumise tõttu jääb seondumissait nukleosoomidest vabaks (Felsenfeld *et al*., 1996). Sellised seondumissaidid tähistavad ligipääsetavaid kromatiini piirkondi. Kromatiini ligipääsetavust on analüüsitud kokku 125 eri raku- ja koeliigis (sh Epstein-Barri viirusega nakatatud B-rakud) (Thurman *et al*., 2012). Leiti, et valdav enamus (97,4%) eelnevalt annoteeritud *cis* mõjuga regulatoorsetest elementidest (sh enhanserid ja insulaatorid) paikneb avatud kromatiini piirkonnas. Samas uuringus leiti ka 42 erineva transkriptsioonifaktori seondumissaitide summeeritud signaal. Kromatiini avatus ühes rakutüübis (K562) ja transkriptsioonifaktorite seondumissaidid on üle genoomi jagunenud väga sarnase mustri järgi (Joonis 1). Kromatiini avatus K562 rakkudes ja transkriptsioonifaktorite seondumissaidid on üle genoomi tugevalt korreleeritud (Pearson r=0,79). See viitab, et kromatiini avatust on võimalik kasutada transkriptsioonifaktorite seondumise hindamiseks, kuigi selle abil ei ole võimalik kindlaks teha, millise konkreetse faktoriga on parasjagu tegu.

Kromatiini avatuse kvantitatiivne lookus (caQTL: *chromatin accessibility quantitative trait locus*) on geneetiline variant, mis mõjutab kromatiini avatuse määra ja seega võib anda infot regulaatorvalkude seondumise kohta. Sellised geneetilised variandid võivad olla aluseks raku fenotüübi kujunemise mehhanismide välja selgitamisele.



Joonis 1. DNase I sekveneerimisel määratud kromatiini avatuse ja valk-DNA interaktsioonide sekveneerimisel (ingl *ChIP-seq*: *Chromatin immunoprecipitation sequencing*) määratud transkriptsioonifaktorite seondumissaitide paiknemine 19. kromosoomi lõigul (Thurman *et al*., 2012).

1.2.2. Kromatiini avatuse ja geenieksпressiooni seos

DNase I sekveneerimine on meetod DNA ligipääsetavuse mõõtmiseks (Song ja Crawford, 2010). DNase I tundlike kromatiini piirkondade ja geenieksпressiooni seost uuriti Epstein-Barri viirusega nakatatud B-rakkudes (LCL: *lymphoblastoid cell line*, n=70) (Degner *et al*., 2012). Ka selles analüüsis leiti, et avatud piirkonnad kattuvad suures ulatuses eelnevalt annoteeritud regulatoorsete aladega. Regulatoorne ala võib sisaldada transkriptsioonifaktori seondumissaiti. Samas uuringus tuvastati kromatiini avatuse QTLde alleelispetsiifiline mõju transkriptsioonifaktorite seondumisele. Erinevus transkriptsioonifaktorite seondumises võib omada mõju geenieksпressioonile.

Sellega kooskõlas leiti, et geneetilised variandid, mis mõjutavad DNase I tundlikkust mõjutavad 16% juhtudest ka vähemalt ühe lähedal asuva geeni ekspressiooni. Samas analüüsis leiti olenevalt analüüsi metoodikast, et 23-55% ekspressiooniga seotud geneetilistest variantidest mõjutavad ka DNase I tundlikkust. See näitab, et kromatiini avatuse varieeruvusel on geenieksпressiooni regulatsioonil märkimisväärne roll.

DNase I sekveneerimine on ATAC sekveneerimise (Buenrostro *et al*., 2015) alternatiiv. Seega oleks mõistlik hüpotees, et ka ATAC sekveneerimisel leitud avatud piirkondadega ("kühmudega") seotud variantide kohta võib leida sarnaseid tulemusi. LCL rakkudes (n=24) uuriti ka ATAC sekveneerimisel kvantifitseeritud kromatiini avatuse QTLe (Kumasaka *et al*., 2016). Leiti, et suur osa kromatiini avatust mõjutavaid geneetilisi variante paikneb selles samas avatud kühmus. Nendest kühmusisestest geneetilistest variantidest suur osa (69%) paiknes mõne transkriptsioonifaktori seondumissaidis. See leid on kooskõlas DNase I analüüsil leituga ja viitab kromatiini avatuse ja geeniekspressiooni märkimisväärsele seosele.

1.3. Splaissimine

1.3.1. Geeniekspressiooni ja splaissimist mõjutavad erinevad geneetilised variandid

LCL rakkudes uuriti erinevate kvantitatiivsete tunnustega seotud *cis* mõjuga geneetilisi variante (Li et al., 2016). Muuhulgas vaadeldi transkriptsioonifaktor CTCF seondumist, DNase I tundlikkust, geeniekspressiooni ja splaissimist. Uuringus leiti, et enamus splaissimisega seotud geneetilised variandid ei mõjuta üldist geeniekspressiooni taset. Teisisõnu need ei mõjuta toodetud RNA hulka, kuid võivad mõjutada lõplikku mRNA järjestust ja seeläbi valkude funktsiooni.

Samal ajal leiti, et nii ekspressiooni kui splaissimisega seotud variandid on ülekattes GWAS (ingl *genome wide association study*) uuringutes haiguste või komplekstunnustega seostatud variantidega. See näitab, et splaissimisel on organismi fenotüübi kujunemisel geeniekspressioonist eraldiseisev mõju.

1.3.2. Alternatiivne splaissimine ja selle seos haigustega

Organismi elutegevuseks vajalike valkude arv on tunduvalt suurem kui erinevate valke kodeerivate geenide arv (Matlin *et al*., 2005). Toodetud valkude mitmekesisust suurendatakse pre-mRNA alternatiivsel splaissimisel. Pre-mRNA töötlemise mehhanismid saab jagada nelja põhilisse klassi: alternatiivne 5' ots, alternatiivne 3' ots, keskmise eksoni kaasamine või kõrvale jätmine ja intronite kaasamine (Nilsen ja Graveley, 2010).

Alternatiivne splaissimine on väga levinud, see toimub vähemalt 70% inimese geenides (Pan *et al*., 2008). Vead splaissimisel võivad kaasa tuua ebanormaalse mRNA ja valgu tootmise. Alternatiivse splaissimise olulise rolli tõttu on splaissimisvigu võimalik seostada mitmete

geneetiliste haiguste tekkega (Wang ja Cooper, 2007). LCL rakkudes leiti, et kuni 50% inimese haigustega seotud mutatsioonidest mõjutavad splaiss-saidi valikut (Li *et al*., 2016).

1.3.3. Kromatiini avatuse ja splaissimise seos

Alternatiivse splaissimise geneetiline kontroll toimib läbi keerulise RNA molekulide ja valkude vaheliste mõjude võrgustiku (Wang ja Burge, 2008). Neid kontrollmehhanisme on nimetatud splaissimise koodiks (Barash *et al*., 2010). Splaissimise koodi osaks on nii transkripti sees paiknevad *cis* mõjuga geneetilised variandid kui ka *trans* mõjuga faktorid, mis transkriptile seonduvad (Wang ja Burge, 2008).

Mitmes uuringus on tuvastatud kromatiini struktuuri seos alternatiivse splaissimisega. Erinevate kudede analüüsis leiti, et DNase I tundlikud piirkonnad ja CTCF seondumissaidid sisaldavad proportsionaalset rohkem alternatiivse splaissimisega seotud geneetilisi variante (Gutierrez-Arcelus *et al*., 2015). Splaissimisega seotud geneetiliste variantide kattumine kromatiini avatusega seotud geneetiliste variantidega leiti ka LCL rakkude analüüsil (Li *et al*., 2016).

1.3.4. Splaissimise regulatsioonimehhanismid

Spaissimisel eksonite kaasamist katalüüsib valguline kompleks splaissosoom (Sharp, 1988). Iga introni otstes leiduvad splaiss-saidis, mille valikust sõltub lõpliku transkripti järjestus (Breathnach ja Chambon, 1981). Leidub nõrgemaid ja tugevamaid splaiss-saite. Tugevamad splaiss-saidid on splaissosoomi poolt kergemini äratuntavad (Matlin *et al*., 2005). Splaiss-saitide valikut kontrollib splaissimise kood. Splaissimisfaktorite seondumine võimendavatele või vaigistavatele järjestustele (Matlin *et al*., 2005) ja RNA polümeraas II aeglustumine (Schor *et al*., 2013) võib kaasa tuua nõrgema splaiss-saidi valiku. Splaissimise regulatsioonis on varem näidatud olulist rolli epigeneetilistel markeritel ja valkudel HP1 ning CTCF (Yearim *et al*., 2015) (Agirre *et al*., 2015) (Shukla *et al*., 2011) (Ruiz-Velasco *et al*., 2017). Kromatiini regulatoorset rolli alternatiivsel splaissimisel (rinna koe rakkudes) uuriti masinõpet kasutades (Agirre *et al*., 2015). Töös kasutati andmeid HP1, CTCF, AGO1 (Argonaut), RNA-polümeraas II seondumisest (ChIPseq) ja histoonide modifikatsioone. Leiti, et lõplikku transkripti kaasatavad eksonid on HP1 ja CTCFga rikastatud. Kõigi nende kromatiini tunnuste arvesse võtmisel õnnestus ennustada kuni 70% alternatiivse splaissimise sündmustest. Kusjuures HP1 ja CTCF olid ühed parimad tunnused, mille põhjal eksonite

kaasamist ennustada. Samas töös leiti RNA polümeraas II ChIPseq signaali seos nii eksonite kaasamise kui kõrvale jätmisega.

Katseliselt leiti metülatsiooni regulatoorne roll alternatiivsel splaissimisel (Yearim *et al*., 2015). HP1 võib olla vahelüliks, mis ühendab metülatsiooni ja splaissimise vahendades transkriptsioonifaktorite seondumist. Leiti, et metülatsioon ja HP1 seondumine võib kaasa tuua nii eksonite kaasamise kui kõrvale jätmise. Samas analüüsis leiti tõendeid, et HP1 mõju splaissimisele sõltub sellest, kas see on seondunud metüleeritud või metüleerimata eksonile. HP1 metüleeritud eksonil on splaissimise võimendaja (ingl *enhancer*) ja metüleerimata eksonil vaigistaja.

Ka CTCF mõju splaissimisele võib olla seotud DNA metüleeritusega. CTCF seondumiskohtade metülatsioon blokeerib faktori seondumist (Hashimoto *et al*., 2017). Leiti, et CTCF võib kontrollida alternatiivset splaissimist läbi RNA polümeraas II aeglustamise (Shukla *et al*., 2011), mis toob kaasa nõrgema splaiss-saidi kasutuse. Ilma CTCF seondumiseta ei toimu RNA polümeraas II aeglustumist ja seeläbi võib metülatsioon vähendada alternatiivsete eksonite kaasamist (Linker *et al*., 2019).

Ruiz-Velasco *et al.* kirjeldasid alternatiivset CTCF toimemehhanismi, kus kaks kromatiiniga seondunud CTCF molekuli ühinevad ja moodustuda kromatiinist aasa, mis lähendab muidu kaugel paiknevaid piirkondi. Analüüsis leiti, et eksoni ja promootori vahel esinevad CTCF aasad korreleeruvad eksonite erineva kasutusega (Ruiz-Velasco *et al*., 2017). Võimalik on erinevate kontrollmehhanismide koosmõju. Linker *et al.* leidsid, et CTCF seondumiskohtade leidumine eksonis on sobiv tunnus eksonite kaasamise ennustamiseks (Linker *et al*., 2019).

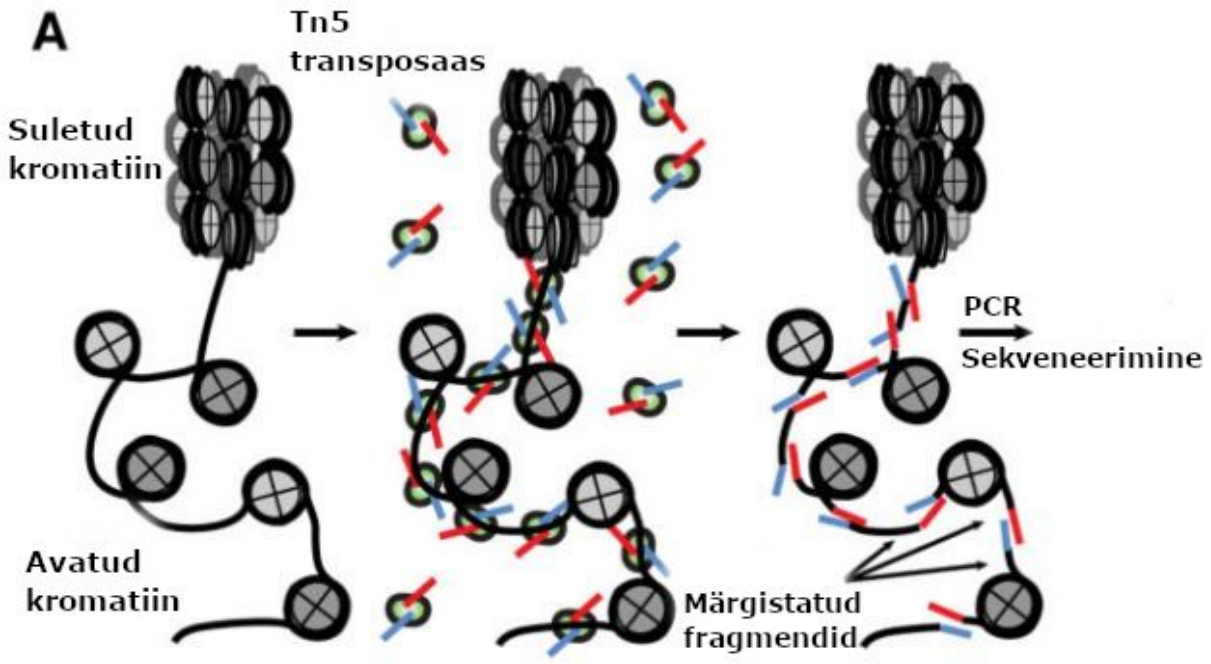
1.4. Mõõtmistehnoloogiad

1.4.1. Kromatiini avatus

DNase-seq ja ATAC-seq on meetodid, millega on võimalik tuvastada avatud kromatiini piirkonnad terve genoomi ulatuses. Mõlemad tehnoloogiad põhinevad sellel, et avatud kromatiin lõigatakse fragmentideks ja fragmendid sekveneeritakse. DNase-seq (ingl *Mapping DNase I hypersensitive sites with high-throughput sequencing*) põhineb endonukleaas DNase I kasutamisel. Nukleosoomidest vabad kromatiini piirkonnad on DNase I lõikavale aktiivsusele tundlikud (Song ja Crawford, 2010).

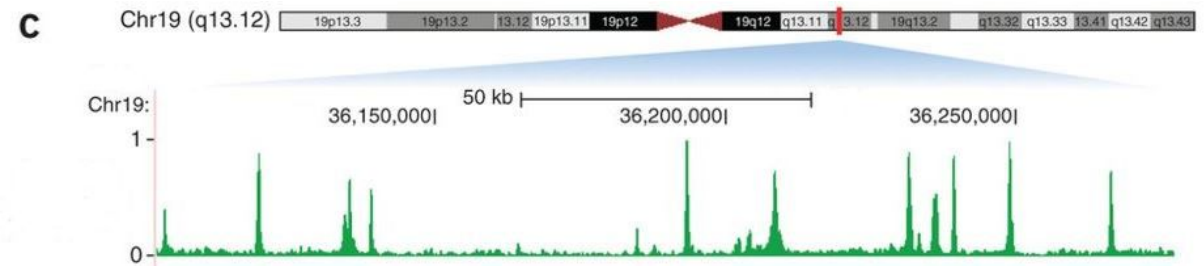
ATAC-seq (ingl *Assay for Transposase Accessible Chromatin with high-throughput sequencing*) põhineb transposaas Tn5 kasutamisel. Tn5 teeb avatud kromatiini lõike ja

ligeerib sekveneerimiseks spetsiifilise adapteri (Buenrostro *et al*., 2015). Mõlema analüüsi puhul järgneb kromatiini lõikamisele amplifikatsioon polümeraasi ahelreaktsioonil ja lugemite sekveneerimine. Kusjuures DNase-seq korral tuleb sekveneerimiseks vajalikud adapterid eraldi sammuna fragmentide külge ligeerida (Song ja Crawford, 2010)(Buenrostro *et al*., 2015). ATAC sekveneerimisel on mitmeid eelised. DNase I sekveneerimiseks vajalik laboriprotokoll on tunduvalt aeganõudvam. ATAC-seq on välja töötatud kiirust silmas pidades. Samuti on vajalik rakkude hulk tunduvalt väiksem. DNase sekveneerimisel kasutatakse standardselt 50 miljonit rakku, ATAC-seq vajab 50 000 rakku (Tsompana ja Buck, 2014).



Joonis 2. Tn5 (rohelised) ja adapterite (punased ja sinised) abil avatud kromatiini tuvastamine ATAC sekveneerimisel (Buenrostro *et al*., 2015).

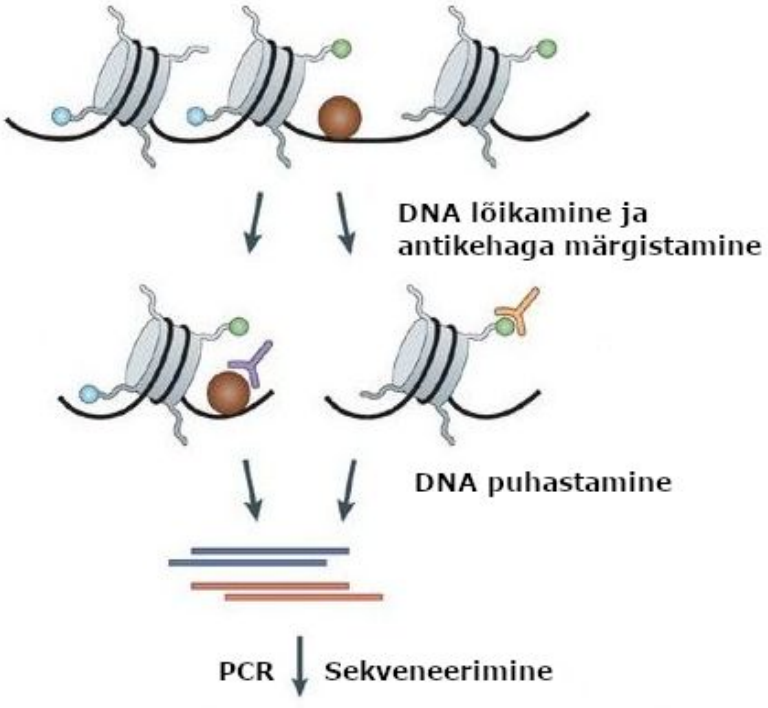
Sekveneerimisele järgneb lugemite joondamine viitegenoomile (ingl *reference genome*). Mõlema meetodi korral on tulemuseks ühe aluspaari täpsuse resolutsiooniga pilt kromatiini avatuse määrast üle genoomi. Selle põhjal on arvutusliku analüüsi abil võimalik tuvastada avatud kromatiini kühmud.



Joonis 3. Näide ATAC sekveneerimise signaalist 19. kromosoomi lõigul (Buenrostro *et al*., 2015).

1.4.2. Transkriptsioonifaktorite seondumine

Transkriptsioonifaktorite seondumiskohtade tuvastamiseks sobiv meetod on valk-DNA interaktsioonide sekveneerimine (ingl *ChIP-seq*: *Chromatin immunoprecipitation sequencing*) (Raha *et al*., 2010). Analüüsis viiakse esmalt läbi DNA ja seostunud valkude ristseostamine (ingl *cross-linking*) formaldehüüdi abil. Sellele järgneb rakkude lüüsimine ja kromatiini lõikamine 200-600 aluspaari pikkusteks fragmentideks. Nende järjestuste peal kasutatakse valgu-spetsiifilist antikeha. Antikehaga märgistatud fragmendid eraldatakse ja neilt eemaldatakse ristseosed (Park, 2009). Sarnaselt kromatiini avatuse analüüsile teostatakse ka ChIP-seq puhul PCR amplifitseerimine ja seejärel sekveneerimine. Kuna on toimunud kromatiini lõikamine juhuslikeks fragmentideks ei võimalda ChIP-seq erinevalt kromatiini avatuse sekveneerimisest ühe aluspaari täpsust resolutsiooni.



Joonis 4. Valk-DNA interaktsioonid ChIP sekveneerimisel (Park, 2009).

Sekveneerimise järgselt joondatakse lugemid viitegenoomile. ChIP-seq lugemid ei kattu täpselt uuritava valgu seondumiskohaga. Lugemid pärinevad loodud fragmentide otstest. Seetõttu moodustuvad mõlemale poole arvatavat seondumissaiti piirkonnad, mida katab rohkem lugemeid. Sellistest algsetest joondatud lugemitest on vaja arvutuslikul analüüsil leida transkriptsioonifaktori seondumisega rikastatud kühmud (Zhang *et al*., 2008).

1.4.3. Transkriptsioon

Geenieksрessiooni taseme kvantifitseerimiseks viiakse läbi RNA sekveneerimine (ingl *RNA sequencing*). RNA transkriptid pöördtranskribeeritakse esmalt komplementaarseks DNAks (cDNA) ja seejärel sekveneeritakse need lugemid. RNA sekveneerimise tulemuseks on kõikide rakus leiduvate transkriptide järjestused ja nende hulk (Wang *et al*., 2009).

2. Eksperimentaalosa

2.1. Töö eesmärgid

Antud töö eesmärgiks on ATAC ja RNA sekveneerimise andmete põhjal leida geneetilisi variante, mis mõjutavad samaaegselt kromatiini avatust ja alternatiivset splaissimist. Faktorite CTCF ja HP1 seondumiskohtade põhjal on eesmärgiks välja selgitada, kas faktoritel on läbi kromatiinile kinnitumise splaissimise regulatsioonis märkimisväärne roll.

Fisheri täpsete testide abil on eesmärgiks kinnitada või ümber lükata järgnevad väited:

1. Kromatiini avatusega seotud lookused on väiksema tõenäosusega seotud splaissimise kui üldise geeniekpressiooni või promootori valikuga.
2. CTCF või HP1 seondumiskohta sisaldavate avatud kromatiini piirkondade hulgas on suurem osakaal splaissimisega seotud piirkondi kui kõigi avatud piirkondade hulgas.
3. Lookused, mis on caQTLd ja CTCF QTLd on suurema tõenäosusega seotud splaissimisega kui kõik caQTLd.

Töös olen keskendunud splaissimise regulatsiooni uurimisele. Üldise geenieksрessiooni, promootori kasutuse ja 3’ otsa valiku andmeid olen uurinud, et võrrelda neid splaissimise analüüsi leidudega.

2.2. Materjal

Kõik töös kasutatavad andmed pärinevad Epstein-Barri viirusega nakatatud B-rakkudest (ingl *LCL*: *lymphoblastoid cell line*).

2.2.1. Kumasaka et al.

Kromatiini avatuse analüüsiks kasutatud rakuliinid on saadud 1000 genoomi projektist (The 1000 Genomes Project Consortium, 2015) ja ATAC-seq viidi läbi Kumasaka ja kaasautorite poolt (Kumasaka *et al*., 2016). Töös kasutan 91 Suurbritannia päritolu indiviidi rakuliini andmeid. DNA eraldati Epstein-Barri viirusega nakatatud lümfoblastidest.

ATAC-seq analüüs viidi läbi sarnaselt varem kirjeldatud metoodikale (Buenrostro *et al*., 2013) Müra vähendamiseks tehtud muudatused metoodikas on kirjeldatud algse artikli lisades (Kumasaka *et al*., 2016). Materjal pärineb 100 000 LCL raku tuumast. Loodud ATAC genoomsetest raamatukogudest filtreeriti välja 120-1000 aluspaari pikkused fragmendid. 75 aluspaari pikkuste lugemite paaris-otsaline (*paired-end*) sekveneerimine viidi läbi HiSeq 2500 (Illumina) masinal. Tulemuseks saadi 892 millionit autosoomsetele kromosoomidele joonduvat lugemit (Kumasaka *et al*., 2016). Enda analüüsi alustasin Kumasaka *et al.* töö tulemusel saadud joondamata ATAC-seq fragmentide fastq failidest (*European Nucleotide Archive* identifikaator ERP011141).

VCF (ingl *variant call format*) on failiformaat genotüübi andmete salvestamiseks. VCF failis on iga lookuse jaoks viitealus (ingl *reference base*) ning iga indiviidi jaoks viitealuste arv igas lookuses. Samade indiviidide genotüübi VCF failid laadisin alla 1000 genoomi projekti veebilehelt (http://www.internationalgenome.org/data, 2018). Iga proovi jaoks on genotüübitud peaaegu 6 miljonit lookust, sealhulgas ühenukleotiidsed polümorfismid (ingl *single nucleotide polymorphism*) ja indelid (ingl *insertion or deletion*).

2.2.2. GEUVADIS

RNA-seq andmed on pärit GEUVADISe projektist (Lappalainen *et al*., 2013). RNA eraldati 465 indiviidi LCL rakuliinidest. Rakuliinid pärinesid viiest erinevast populatsioonist: CEPH (*Utah Residents with Northern and Western European Ancestry*), Soome, Suurbritannia, Toskaana, Yoruba. Nendest 445 proovi jaoks on genotüübi andmed kättesaadavad *1000 genoomi* projekti kolmandast faasist (The 1000 Genomes Project Consortium, 2015). Käesolevas töös on kasutatud nendest 358 Euroopa päritolu indiviidi andmeid.

2.2.3. HP1 ja CTCF ChIP-seq

Mõlema tranksriptsioonifaktori puhul on analüüs sooritatud GM12878 LCL rakuliini peal, mis pärineb CEPH populatsioonist. Transkriptsioonifaktorite CTCF ja HP1 seondumissaidid leiti ENCODE konsortsiumi poolt ChIP sekveneerimisel. Seondumissaitide failid laadisin alla ENCODE portaalist (Davis *et al*., 2018) (https://www.encodeproject.org/, 2018). CTCF ja HP1 andmetele vastavad identifikaatorid ENCFF960ZGP ja ENCFF417SVR. Jätsin kõrvale seondumiskohad X kromosoomil (Tabel 1).

Tabel 1. ENCODE andmestiku transkriptsioonifaktorite seondumiskohad

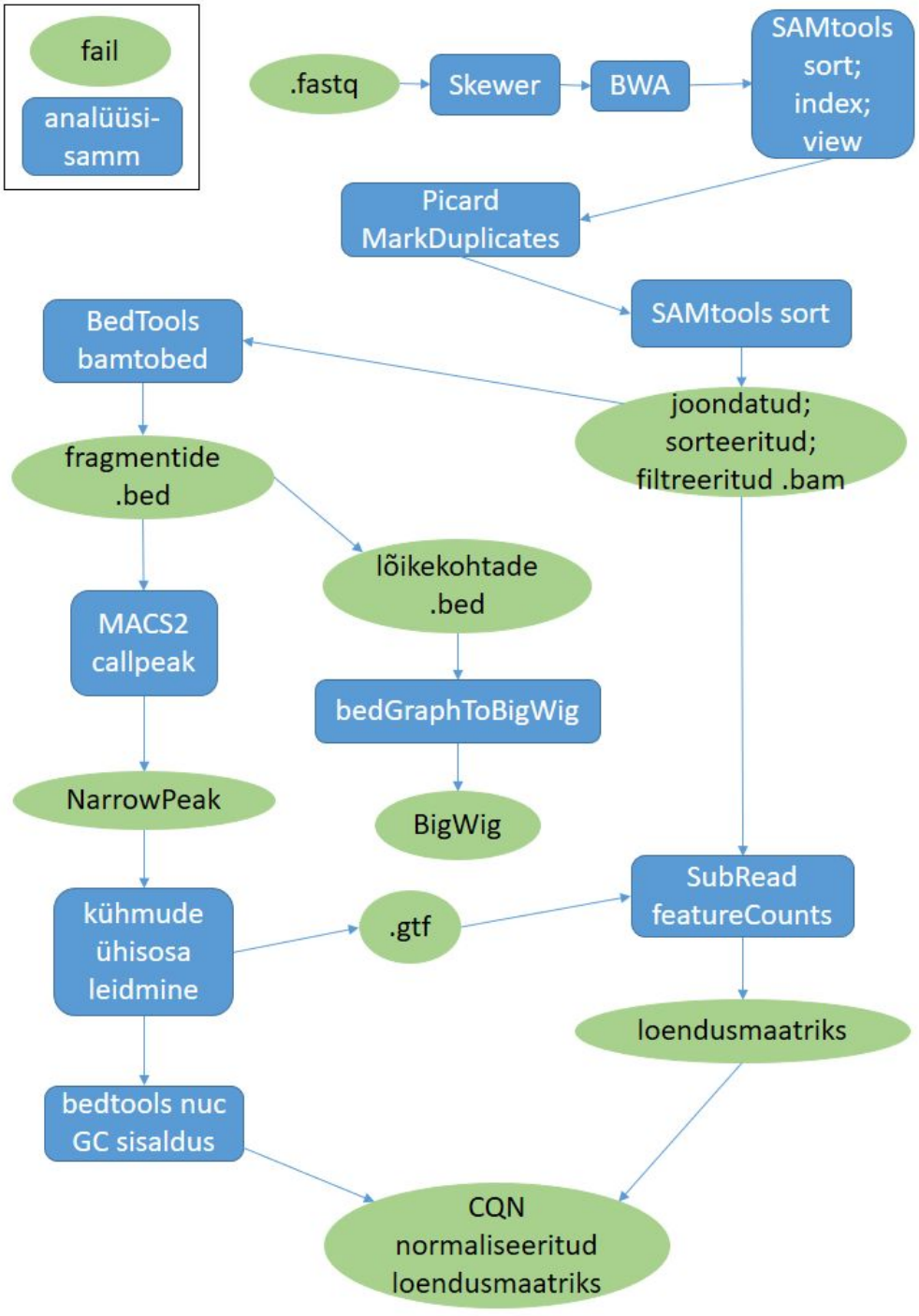
	seondumiskohti kokku	seondumiskohti somaatilistel kromosoomidel
HP1	10 855	10 610
CTCF	43 865	42 397

2.2.4. CTCF QTLid

CTCF QTLid on geneetilised variandid, millel on statistiline seos CTCF seondumisega. ChIP-seq on sooritatud 51 CEPH populatsioonist pärit indiviidi LCL rakkudes (Ding *et al*., 2014). CTCF ChIP-seq andmete põhjal tuvastati MACS tööriista abil CTCF seondumissaidid (Zhang et al., 2008). QTLtools on molekulaarsete kvantitatiivsete tunnuste analüüsimiseks loodud käsurea tööriist (Delaneau *et al*., 2017). Kasutades CTCF seondumissaite ja genotüübi andmeid sooritati QTLtools abil juhuslikud genotüüpide permutatsioonid. Nii leiti geneetilised variandid, millel on statistiliselt oluline seos CTCF seondumisega. Need on kvantitatiivse tunnuse lookused (QTLd). Enda analüüsi alustasin QTLtools väljundfailist, mis olid loodud Yurii Toma magistritöö raames (Toma, 2018). Kasutasin nii CTCF seondumissaite kui leitud QTLe.

2.3. Meetodid

2.3.1. DNA ligipääsetavuse analüüs



Joonis 5. DNA ligipääsetavuse sekveneerimisandmete analüüsi töövoog.

Kromatiini ligipääsetavuse analüüsi teostasin 91 indiviidi sekveneerimisandmetel. Töötlemata kromatiini avatuse andmed olid fastq formaadis ning ühele indiviidile vastas mitu eri faili. Koondasin iga indiviidi andmed eraldi faili kasutades UNIX käsklusi. Eemaldasin

sekveneerimisega kaasnevad vöötkoodid ja adapterid skewer tööriista abil (Jiang *et al*., 2014). Joondasin viitegenoomile GRCh38 kasutades BWA (versioon 0.7.12) tööriista (Li ja Durbin, 2009). Lugemid sorteerisin genoomse koordinaadi järgi ja indekseerisin kasutades käsurea tööriista SAMtools (versioon 1.3) (Li *et al*., 2009). SAMtoolsi kasutasin ka mitokondriaalse DNA lugemite eemaldamiseks ning failide päise asendamiseks. Loodud BAM formaadis failidest eemaldasin korduvad lugemid kasutades MarkDuplicates tööriista Picard paketist (http://broadinstitute.github.io/picard/, 2018). Korduvad lugemid on PCR amplifitseerimisel tekkinud artefaktid.

Järgmisena sorteerisin failid SAMtools abil lugeminime järgi. Järjestatud BAM formaadis faili konverteerisin BEDTools (versioon 2.24) (Quinlan ja Hall, 2010) tööriista abil BED formaati ning bedGraphToBigWig (https://www.encodeproject.org/software/bedgraphtobigwig/, 2018) tööriista abil BigWig formaati. ATAC sekveneerimisel on DNA lõigatud transposaas Tn5 poolt (Buenrostro *et al*., 2015). Edasi leidsin lugemite põhjal transposaas Tn5 lõikekohad. Lõikekohtade BAM failide põhjal leidsin kromatiini avatusele vastavad narrowPeak failid, kasutades tööriista MACS (versioon 2.1.0 parameetritega --shift 25 --extsize 50 -q 0.01 --*t size*: 75) (Zhang *et al*., 2008). Transposaasi lõikekohtade põhjal soovisin leida laiemad piirkonnad, kus kromatiin on avatud ehk avatuse "kühmud". Kühmude leidmiseks on *shift* ja *extsize* parameetrite põhjal genereeritud pseudolugemid, mis ulatuvad lõikekohast mõlemas suunas 25 aluspaari kaugusele. Lugemitega kattuvate alade põhjal tagastas tööriist piirkonnad, mille p-väärtus on mitmese testimise vastu korrigeeritud FDR (*false discovery rate*) meetodil. Edaspidi vaatlen ainult statistiliselt olulisi (FDR < 0.01) kühme.

MACS (ingl *Model-based analysis of ChIP-Seq*) käsurea tööriist on loodud ChIP sekveneerimisel valkude seondumiskohtade leidmiseks. Sisendiks on ette nähtud fragmendid, mis katavad huvipakkuva fenotüübiga kromatiini ala. ATAC sekveneerimisel tuvastatakse üksikud Tn5 lõikekohad, mitte avatud lõigud. Seetõttu tuleb MACS parameetreid kohandada ja pseudolugemid kunstlikult genereerida.

Kvaliteedikontrolli eesmärgil leidsin fragmentide arvu kromosoomi kohta ning fragmentide pikkuse jaotuse kasutades SAMtools ja UNIX käsklusi. Igale indiviidile vastavate kühmude põhjal leidsin kühmude ühisosa üle kõigi proovide. Need on kõik kromatiini piirkonnad, kus ATAC kühm esineb vähemalt kolmes eri indiviidis. Leidsin 20 kuni 1000 aluspaari pikkuste fragmentide arvu iga indiviidi jaoks igas ühendatud kromatiini avatuse piirkonnas kasutades

SubRead (versioon 1.6.2) paketi tööriista featureCounts (Liao *et al*., 2014). Eemaldasin analüüsist kühmud, millest 50 000 aluspaari kaugusel ei ole genotüübitud ühtegi lookust.

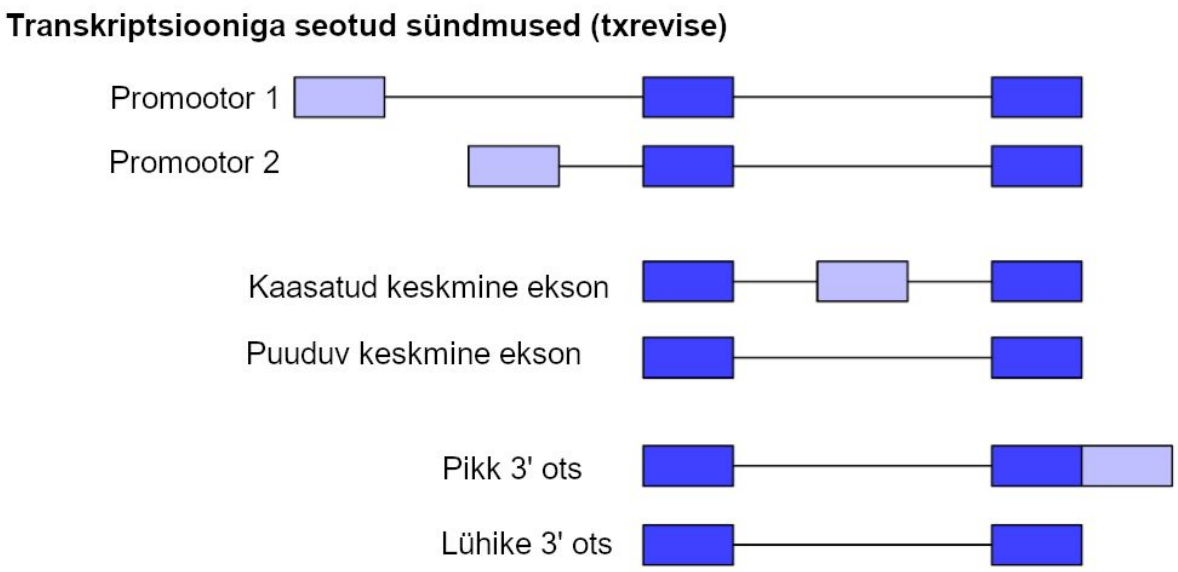
BEDTools abil leidsin kühmude piirkonnale vastava guaniini ja tsütosiini sisalduse protsendi viitegenoomis. Guaniini ja tsütosiini sisaldust kasutasin hiljem normaliseerimisel. Ühendasin kõikide indiviidide fragmentide loendusandmed ühte maatriksisse. Loendusandmed normaliseerisin kasutades Bioconductor (versioon 3.8) CQN (Hansen *et al*., 2012) paketti. Selle töö tulemuseks on CQN (ingl *conditional quantile normalization*) normaliseeritud DNA ligipääsetavuse loendusmaatriks, kus ridadeks huvipakkuvad kühmud ja veergudeks 91 indiviidi.

Koondamaks algsed eraldi loendusandmete failid ühisesse normaliseeritud maatriksisse tuli läbida iga proovi jaoks rohkem kui 10 sammu, kus kasutasin R või Python skripte või käsurea tööriistu. Kuna selliste tehtavate sammude arv on suurusjärgus 1000, tuli leida viis selle töö automatiseerimiseks. Analüüsi töövoo juhtimiseks kasutasin Snakemake tööriista (Köster ja Rahmann, 2012). Snakemake võimaldas vajalikud sammud järjestada ning iga faili jaoks üldistada. Kasutatud Snakemake skript põhineb Kaur Alasoo tööl (https://github.com/kauralasoo/Blood_ATAC/blob/master/Snakefile, 2018).

2.3.2. Geenieksпressiooni ja splaissimise analüüs

Lappalainen *et al.* loodud RNA-seq andmete edasise analüüsi teostas Kaur Alasoo. RNA-seq andmete põhjal on transkriptide varieeruvus kvantifitseeritud neljal eri viisil. Eraldi tunnusena on käsitletud promootori kasutust, splaissimist, transkripti 3’ otsa valikut ja geeniga kattuvate lugemite arvu. Varieeruvus promootorite, transkripti 3’ otsa valikus ja splaissimises mõõdeti txrevise meetodi abil (Alasoo *et al*., 2018). Geenidega kattuvate lugemite loendamiseks on kasutatud featureCounts programmi (Liao *et al*., 2014).

Ensemblist on võetud eelnevalt annoteeritud transkriptid. Iga geeni jaoks on üldine geenieksпressioon kvantifitseeritud nende transkriptidega kattuvate lugemite arvu põhjal. Txrevise meetodi korral on Ensembl annotatsiooni põhjal leitud iga geeni jaoks eksonid, mis leiduvad igas transkriptis (*constitutive exon*). Unikaalsetele eksonitele (need, mis ei leidu igas transkriptis) on määratud vastavalt paiknemisele tüübiks alternatiivne promootor, keskmine ekson või alternatiivne 3’ ots.



Joonis 6. Transkriptsiooniga seotud sündmuste kvantifitseerimine (Alasoo *et al.*, 2018). Keskmise eksoni kaasamine või puudumine tähistab käesolevas töös uuritud splaissimist.

2.3.3. Geneetiliste seoste leidmine

Leidmaks seoseid molekulaarse fenotüübi tunnuste ja genotüübi vahel on kasutatud QTLtools tööriista (Delaneau *et al.*, 2017) (versioon 1.1) (parameetrid: cis --permute 10 000 --window 100 000). Siin on nullhüpoteesiks, et fenotüüp ei ole lineaarselt sõltuv genotüübist. Testin alternatiivset hüpoteesi, et fenotüüp on lineaarselt sõltuv ühe ümbritseva lookuse genotüübist. Kromatiini ligipääsetavuse jaoks viisin QTL analüüsi läbi ise. RNA ja CTCF andmete analüüsi alustasin QTLtools väljundfailidest. Iga molekulaarse fenotüübi tunnuse (so kromatiini ligipääsetavus, transkripti kasutus, CTCF seondumine) jaoks on läbi viidud samalaadne analüüs.

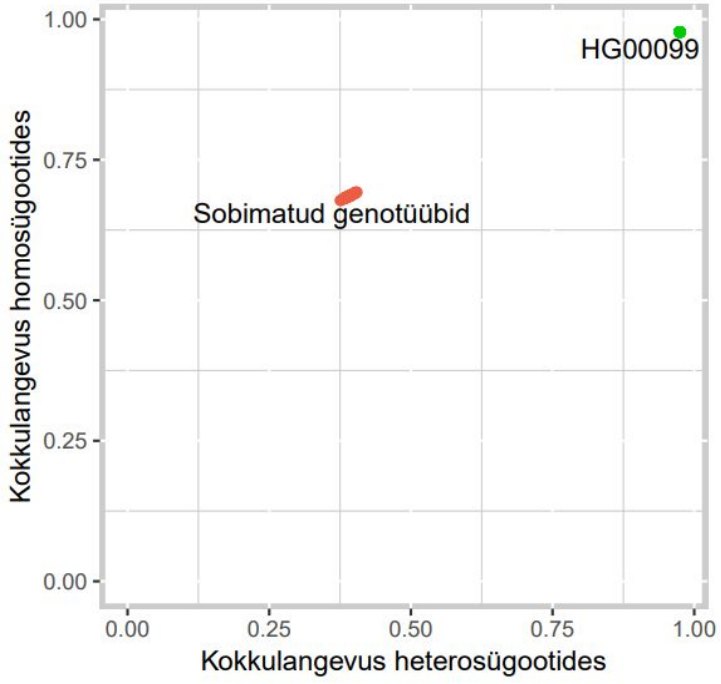
QTLtools jooksutamisel sobitatakse kõigi fenotüüpi ümbritsevate genotüübitud lookuste jaoks lineaarse regressiooni mudel. Muldel sobitatakse iga piirkonnas asuva geneetilise variandi jaoks eraldi, so hinnatakse iga geneetilise variandi mõju tunnusele sõltumata teistest samas piirkonnas asuvatest variantidest. Selle regressioonisirge tõus on teststatistik, mille põhjal arvutatakse mudeli p-väärtus. P-väärtus näitab, kui tõenäoline oleks näha sama tugevat või tugevamat seost genotüübi ja fenotüübi vahel, kui kehtiks nullhüpotees (seost ei ole). Kui p-väärtus on piisavalt väike, siis saame lükata ümber nullhüpoteesi ning järeldada, et genotüübi ja tunnuse vahel on seos.

Iga fenotüübi tunnuse jaoks testitakse seost kõigi ümbritsevate geneetiliste variantidega ja valitakse välja variant, millel on kõige tugevam seos fenotüübiga (kõige väiksem p-väärtus).

QTLtools kasutab permutatsioone, et hinnata testitud sõltumatute genotüüpide arvu. QTLtools väljundis on p-väärtused korrigeeritud iga fenotüübi jaoks testitud sõltumatute genotüüpide arvu suhtes. Lisaks sellele tuleb arvestada, et üle genoomi on testitud palju fenotüüpe. P-väärtuse üle genoomi vaadeldud fenotüüpide arvu vastu korrigeerimiseks kasutasin R stats paketi meetodit p.adjust parameetriga FDR (ingl *false discovery rate*). Edasi analüüsisin ainult statistiliselt olulisi seoseid (FDR < 0,1). Edasise analüüsi jaoks oli vaja QTL failid kromosoomide järgi indekseerida, selleks kasutasin SAMtools tabix funktsionaalsust.

2.3.4. Kromatiini ligipääsetavuse QTLid

Kromatiini ligipääsetavuse analüüsis vaatlesin fenotüübina kromatiini avatuse kühme, nendega seotud lookused on caQTLd (ingl *chromatin accessibility quantitative trait locus*). QTLde leidmiseks kasutasin genotüübi andmeid ja kromatiini avatuse loendusmaatriksit. Kontrollimaks, et ATAC sekveneerimise proovide identifikaatorite märkimisel pole tehtud viga kasutasin QTLtools tööriista mbv funktsionaalsust (Fort et al., 2017). Leidsin iga ATAC sekveneerimise proovi lugemite jaoks, kui hästi need on kooskõlas genotüüpidega (Joonis 7). Genotüübi ja ATAC sekveneerimise identifikaatorid sobitasin juhul kui esines suur kokkulangevus (heterosügootides ja homosügootides > 0,95). Ühtegi valesti identifitseeritud proovi ei tuvastanud.



Joonis 7. ATAC sekveneerimise proovi ERS798625 kokkulangevus genotüüpidega.

Edasises QTL analüüsis üritan leida juhtumeid, kus üks geneetiline variant mõjutab kromatiini ligipääsetavust. Ühe geneetilise variandi mõju kõigile kühmudele üle genoomi on tõenäoliselt väga väike. Kromatiini ligipääsetavust üle kogu genoomi ja kõigi proovide mõjutavad ka suuremad süstemaatilised mõjud (näiteks sekveneerimise kuupäev ja katsele kulunud aeg). Soovin tuvastada varieeruvust erinevate indiviidide vahel iga konkreetse kühmu puhul. Seetõttu tuleks suurt hulka indiviide ja kühme korraga mõjutavad tegurid välja filtreerida. Selleks on QTLtools analüüsis võimalik kasutada kovariaate.

QTLtools tööriista (versioon 1.1) (parameetriga pca) (Delaneau *et al*., 2017) abil sooritasin kromatiini avatuse loendusmaatriksi ja genotüübi andmete jaoks peakomponentanalüüsi (ingl PCA: *principal component analysis*). Sellele järgnes QTLtools *cis* parameetriga genotüüpide juhuslik permuteerimine ja juhtvariantide p-väärtuste korrigeerimine. Tabel 2 annab ülevaate, kuidas QTL analüüsi parameetrid mõjutasid leitud statistiliselt oluliste caQTLde arvu.

Tabel 2. Statistiliselt oluliste kromatiini avatuse QTLde arv sõltuvalt kasutatud parameetritest. Kovariaatidena kasutasin võrdse arvu kromatiini avatuse loendusmaatriksi ja genotüübi maatriksi komponente.

caQTL arv (FDR < 0,1)	normaliseerimis-meetod	akna laius (bp)	kovariaatide arv	permutatsioonide arv
32 827	CQN	10 000	10	10 000
32 819	FPKM	10 000	10	10 000
25 872	CQN	100 000	10	10 000
25 890	FPKM	100 000	10	10 000
33 718	CQN	10 000	10	100
32 713	CQN	10 000	20	100
25 711	CQN	10 000	40	100
26 719	CQN	100 000	10	100
25 704	CQN	100 000	20	100
19 007	CQN	100 000	40	100

CQN ja FPKM (ingl *fragments per kilobase million*) normaliseerimismeetodite vahel märkimiväärset erinevust ei esinenud. Üldiselt toob väiksem permutatsioonide ja kovariaatide arv kaasa suurema hulga valepositiivseid leide. Edasises analüüsis kasutasin kromatiini avatuse QTLe, mille leidmiseks oli sooritatud 10 000 permutatsiooni, kovariaatidena kasutatud 5 kromatiini avatuse ja 5 genotüübi maatriksi PCA komponenti, 100 000 aluspaari laiust akent kühmu ümbruses ja CQN normaliseerimist (Tabel 2 rida 3).

2.3.5. RNA QTL

RNA QTL analüüsis vaadeldi fenotüübi tunnusena geeniekspressiooni ja splaissimise analüüsis mõõdetud nelja tunnust. Fenotüübi varieeruvusega seotud lookuste (QTLde) leidmiseks kasutati QTLtools tööriista (parameetriga cis) (Delaneau *et al*., 2017). Enda analüüsi alustasin QTLtools väljundfailidest, mille leidmiseks oli sooritatud 10 000 juhuslikku genotüüpide permutatsiooni. QTLtools väljundfailis on iga geeni iga alternatiivse transkripti kohta kokku 19 tunnust. Iga alternatiivse transkripti jaoks on olemas fenotüübi ID ja paiknemine genoomil. Edasises analüüsis kasutasin ainult statistiliselt olulisi QTLe (FDR < 0,1). Vaatlen eraldi promootori kasutusega, alternatiivse splaissimise ehk keskmise eksoni valiku (sQTL), ekspressiooni ja 3’ otsa valikuga seotud lookuseid.

2.3.6. CTCF QTL

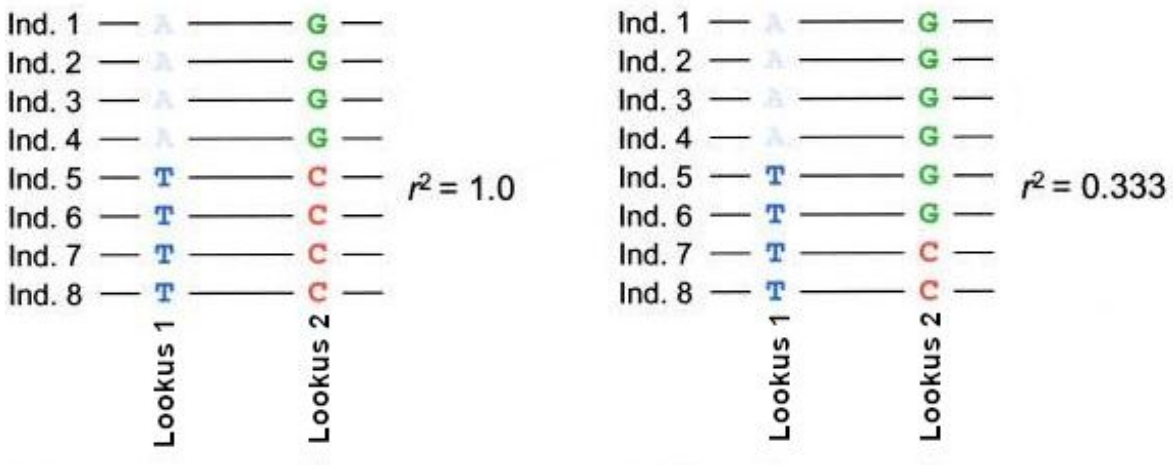
CTCF QTL on geneetiline variant, mille varieeruvusel esineb statistiline seos transkriptsioonifaktor CTCF seondumisega. Yurii Toma leidis somaatilistel kromosoomidel kokku 49 698 CTCF seondumissaiti, millest 1112 puhul leidus statistiliselt oluline (FDR < 10%) QTL. Enda analüüsis leidsin, et 73% siin leitud CTCF seondumiskohtade jaoks leidub ENCODE andmestikus kattuv CTCF seondumiskoht (kattumine vähemalt 50% seondumiskoha laiusest).

2.3.7. Geneetiliste seoste ülekate

Geenide aheldatuse tõttu on samal kromosoomil esinevate lähestiku paiknevate geneetiliste variantide pärandumine korreleeritud. LD (ingl *linkage disequilibrium*) on kahes (või mitmes) lookuses esinevate alleelide seotud pärandumine. LD määra mõjutavad nii demograafilised faktorid kui ka rekombinatsioonisagedus (Pritchard and Przeworski, 2001).

Käesoleva töö eesmärgiks on leida geneetilisi variante, mis mõjutavad üheaegselt kromatiini ligipääsetavust ja alternatiivset splaissimist või CTCF seondumist. QTLtools väljundis on iga

fenotüübi tunnuse jaoks leitud vähima p-väärtusega variant (so juhtvariant). Fenotüüpi tegelikult mõjutav geneetiline variant (so põhjuslik variant) võib olla QTL analüüsis leitud juhtvariant või mõni sellega LDs olev variant. Ka fenotüübid, mille juhtvariandid on LDs omavad suure tõenäosusega sama põhjuslikku varianti. Kahe geneetilise variandi aheldatuse määramiseks kasutasin kriteeriumina väärtust r^2. Väärtuse r^2 leidmiseks võtsin kahe geneetilise variandi genotüübid üle kõigi indiviidide ja leidsin nende Pearsoni korrelatsiooni ruudu (Joonis 8).



Joonis 8. Kahe lookuse aheldatuse määramine 8 indiviidi (Ind. 1-8) genotüüpide põhjal (Gaut ja Long, 2003).

Lihtsaim lähenemine leidmaks geneetilisi variante, mis mõjutavad nii kromatiini avatust kui ka splaissimist oleks leida kõik kromatiini kühmude ja splaissimissündmuste paarid, mille juhtvariandid kattuvad. Aheldatuse mõõt r^2 võimaldab tuvastada rohkem fenotüübi tunnuste paare, mille varieeruvust mõjutab suure tõenäosusega sama geneetiline variant. Aheldatuse analüüsis kasutasin kahte eri lävendit ($r^2 > 0{,}8$ ja $r^2 > 0{,}9$). Aheldatud juhtvariantide paarid leidsin esmalt kromatiini avatuse ja kõigi transkriptsiooniga seotud sündmuste jaoks (geenieksрessioon, promootori kasutus, splaissimine, 3’ otsa valik).

Edasi kasutasin transkriptsioonifaktorite CTCF ja HP1 seondumiskohti. Leidsin avatud kromatiini piirkonnad, mis sisaldavad transkriptsioonifaktori seondumiskohta. Aheldatuse ja seondumiskohtadega kattumise alusel jagasin kromatiini kühmud (iga transkriptsiooniga seotud sündmuse jaoks) nelja klassi. Edasi uurisin, kas transkriptsioonifaktori seondumiskohta sisaldavate kühmude juhtvariandid on suurema tõenäosusega aheldatud splaissimisega seotud variantidega. Selle analüüsi eesmärgiks oli välja selgitada, kas CTCF või

HP1 seondumine on oluline splaissimise kontrollmehhanism. Aheldatud juhtvariantide analüüsi viisin läbi ka Yurii Toma magistritööst pärit CTCF seondumiskohtade jaoks.

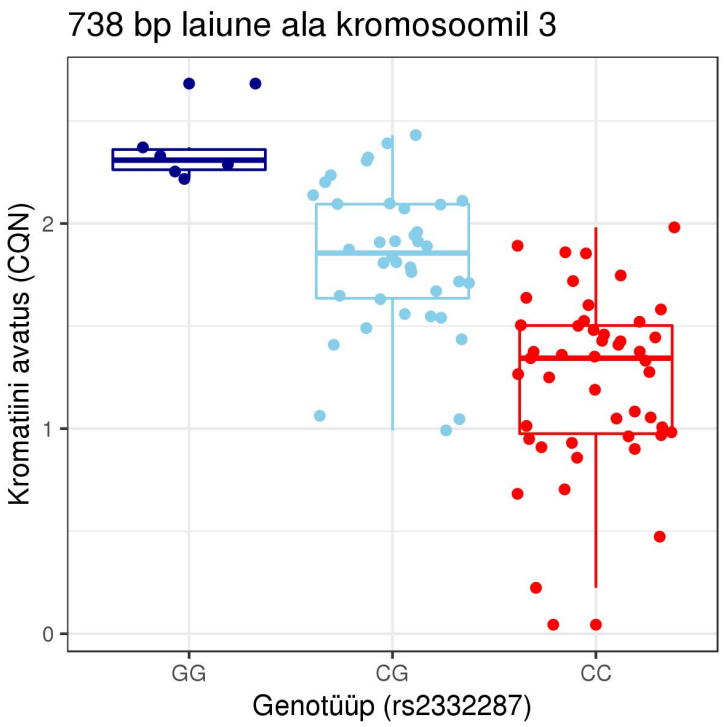
2.4. Tulemused

2.4.1. Seosed kromatiini ligipääsetavuse ja transkriptsiooni vahel

Konstrueerisin ATAC sekveneerimise andmetest 307 517 kromatiini avatuse kühmu, mis esinesid vähemalt kolmes eri indiviidis. Nendest 18 103 jäid edasisest analüüsist kõrvale, sest 100 000 aluspaari kaugusel ei olnud ühtegi genotüübitud lookust. Alles jäänud 289 414 kühmust 8,9% (25 872 kühmu) juhtvariant oli statistiliselt oluline (FDR < 0,1). Statistiliselt olulise juhtvariandiga avatud kromatiini piirkonnad kattusid valdavas enamuses (98%) eelnevalt annoteeritud regulatoorsete aladega (annotatsioonid Ernst *et al.*, 2011). DNase sekveneerimise analüüsil leiti sarnane tulemus (Degner *et al*., 2012). See annab alust väita, et ATAC sekveneerimine on DNase sekveneerimise sobiv alternatiiv.

Varem on leitud, et enamus kromatiini avatust mõjutavaid variante paiknevad avatud piirkonna sees või vahetult selle ümbruses (Degner *et al*., 2012). 100 000 aluspaari ulatuses leidsin 25 872 statistiliselt olulist caQTLi. Leidsin, et kolmandik (32,6% so 8443 kühmu) variantidest, mis mõjutavad kromatiini avatust paiknevad avatud piirkonnast vähem kui 1000 aluspaari kaugusel.

QTLtools anaüüsis sobitatakse lineaarse regressiooni mudelid, seega peaks kromatiini avatus leitud caQTLdes olema genotüübist lineaarses sõltuvuses. Vähimate p-väärtustega QTLde jaoks koostasin selle kontrollimiseks graafikud. Iga lisanduv tsütosiin rs2332287 positsioonil vähendab kromatiini avatust 3. kromosoomil asuva kromatiini avatuse kühmu juures (Joonis 9).



Joonis 9. Juhtvariandi genotüübist sõltuv normaliseeritud kromatiini avatus ühe ATAC kühmu juures. Iga lisanduv tsütosiin juhtvariandi positsioonil vähendab kromatiini avatust.

Tabel 3 annab ülevaate transkriptsiooniga seotud statistiliselt olulistest QTLidest. LCL rakkude DNase sekveneerimise analüüsis leiti olenevalt metoodikast, et 23-55% geenieksprессiooni mõjutavatest lookustest on seotud ka kromatiini avatusega (Degner *et al*., 2012). Käesolevas töös arvestasin, et geen ja kromatiini avatuse kühm on seotud kui neil on ühine juhtvariant või nende juhtvariandid on aheldatud. Leidsin, et 20% geenieksprессioon, 17% promootori, 13% splaissimisega ja 3’ otsa valikuga seotud lookustest mõjutavad suure tõenäosusega ka kromatiini avatust (r^2 > 0,8) (Tabel 3).

Tabel 3. Erinevat tüüpi RNA sekveneerimsandmete töötlusel leitud QTLde jaotus

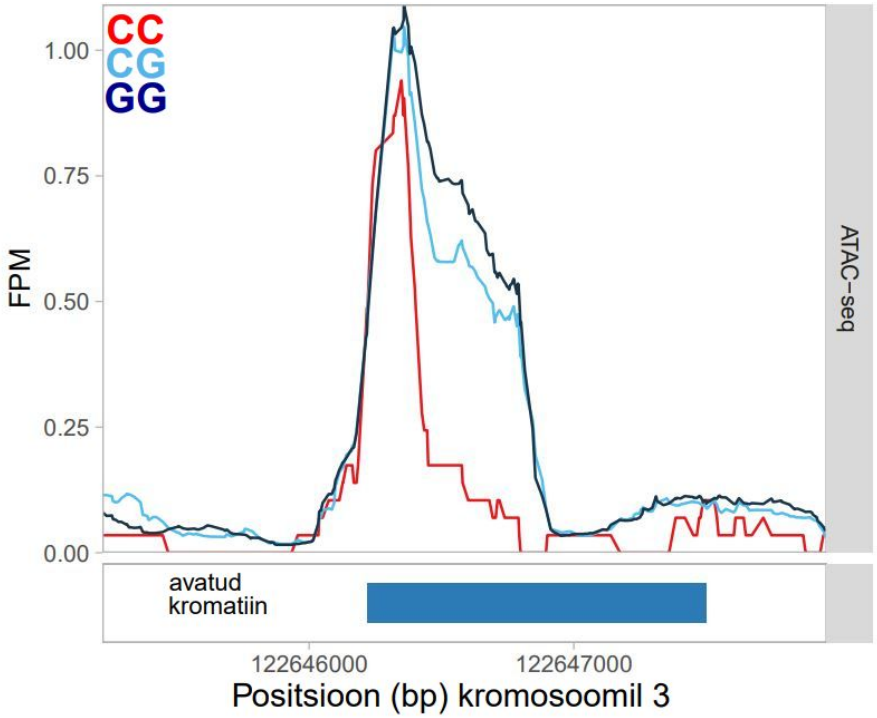
QTL tüüp	Geenide arv kokku	Kromatiini avatusega seotud geenide arv (r^2 > 0.8)	Kromatiini avatusega seotud geenide arv (r^2 > 0.9)
Ekspressioon	5957	1164 (19,5%)	914 (15,3%)
Promootori kasutus	1200	198 (16,5%)	136 (11,3%)
Splaissimine	2716	344 (12,7%)	245 (9,0%)
3’ otsa kasutus	2065	269 (13,0%)	178 (8,6%)

Võrreldes DNase-seq analüüsiga leidsin, et geeniekspressiooniga seotud lookuste hulgas on väiksem osakaal lookuseid, mis on seotud ka kromatiini avatusega. DNase sekveneerimise artiklis ei ole ülekatte arvestamiseks kasutatud r^2 meetodit. Metodoloogiline erinevus võib selgitada kromatiini avatusega seotud ekspressiooni QTLde väiksemat osakaalu.

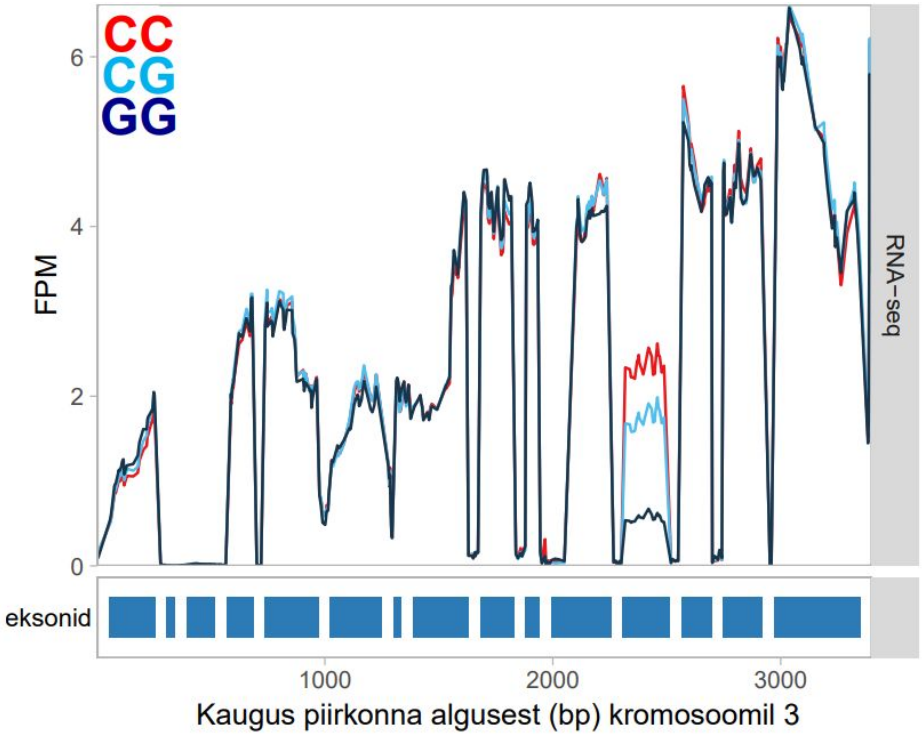
Geeniekpressiooni ja promootori kasutuse QTLde hulgas on suurem osakaal lookuseid, mis on seotud ka kromatiini avatusega kui splaissimise või 3' otsa QTLde hulgas (Fisheri täpse testi p-väärtus < 0,05). See tulemus on ootuspärane. Ekspressioon ja promootori valik toimuvad ajaliselt enne splaissimist. Seega on splaissimise varieeruvuse väiksem seos kromatiini avatusega kooskõlas splaissimise regulatsioonimehhanismidega. Splaiss-saidid paiknevad RNAs, kuid lõplik promootori valik toimub DNA järjestuse tasemel. Ka makrofaagide RNA ja ATAC-seq analüüsis leiti, et promootori valiku QTLde hulgas on suurem osakaal lookuseid, mis on seotud kromatiini avatusega kui splaissimise või 3' otsa QTLde hulgas (Alasoo *et al*., 2018). See tulemus on kooskõlas ka LCL rakkude DNase-seq analüüsiga. Analüüsis leiti, et geeniekspressiooni ja splaissimist mõjutavad erinevad geneetilised variandid (Li et al., 2016).

Analüüsi tulemusel õnnestus leida kühmude ja geenide paare, mille puhul kromatiini avatust ja splaissimist mõjutab suure tõenäosusega sama põhjuslik variant. Seose kontrollimiseks koostasin selliste kühmude ja geenide paaride jaoks joonised, mis kujutavad kattuvust ATAC ja RNA sekveneerimise fragmentidega. Selleks kasutasin BigWig faile, milles on genoomi kattuvus sekveneerimisfragmentidega ühenukleotiidse täpsusega. Joonistel 10 ja 11 kujutan sekveneerimisfragmentide arvu, mis on normaliseeritud sekveneerimiskorduste (ingl *sequencing depth*) järgi kasutades FPM (ingl *fragments per million*) normaliseerimist. Joonistel on kujutatud sekveerimisfragmentide arv sõltuvalt kühmu juhtvariandi (so caQTL) genotüübist (Joonised 10 ja 11). Jooniste tegemiseks kasutasin R paketti wiggleplotr (http://bioconductor.org/packages/release/bioc/html/wiggleplotr.html, 2019).

Väikse p-väärtuse ja suure efektisuurusega caQTL kühmude puhul võib näha selget erinevust kromatiini avatuses erinevate juhtvariandi genotüüpide puhul (Joonis 10). Joonis 10 kinnitab joonisel 9 esitatud tulemust. Iga lisanduv tsütosiin (rs2332287 positsioonil) vähendab kromatiini avatust 3. kromosoomil paikneva kromatiini kühmu ulatuses.

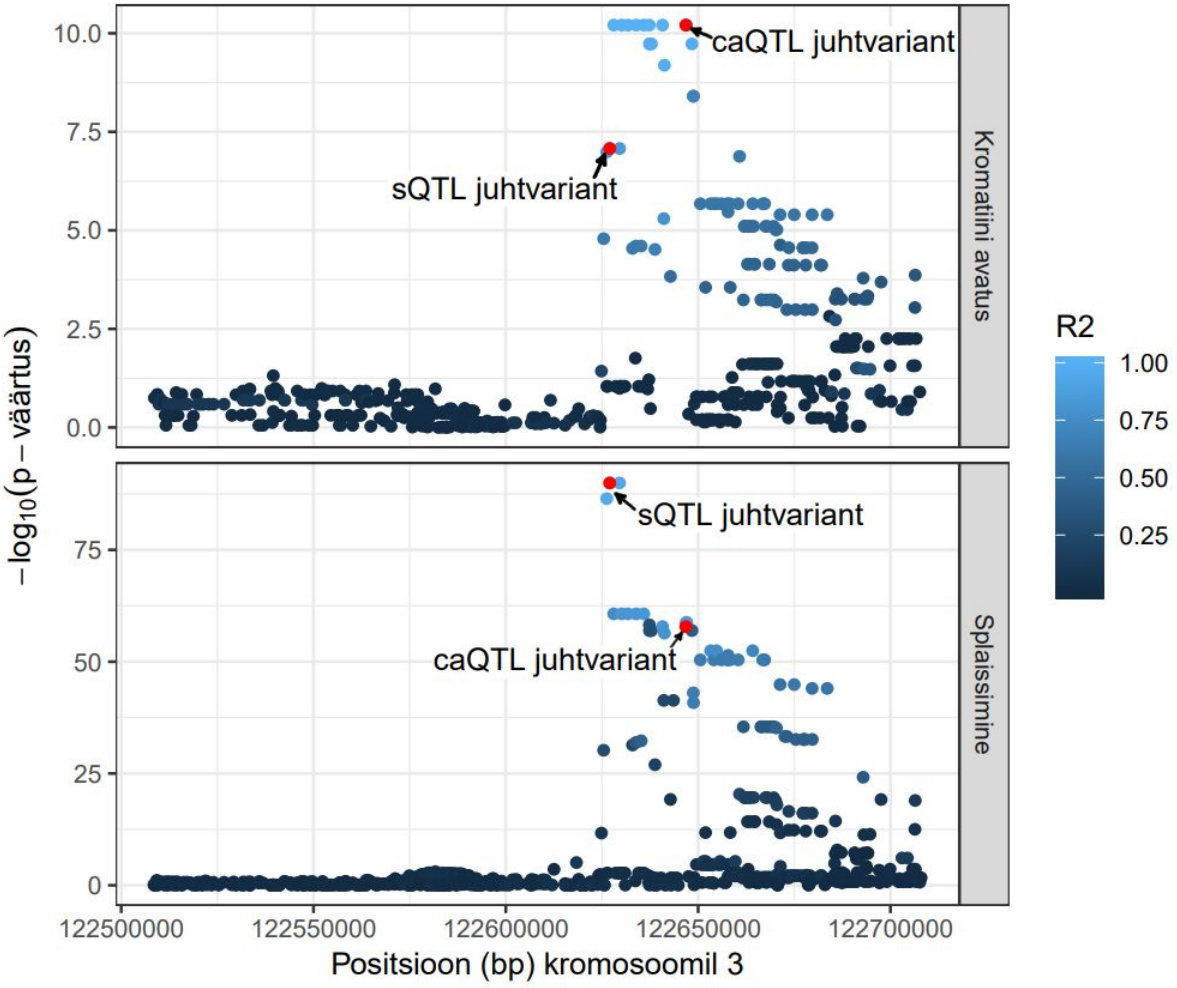


Joonis 10. ATAC sekveneerimise fragmentide arv sõltuvalt kühmu juhtvariandi (rs2332287) genotüübist. Iga lisanduv tsütosiin juhtvariandi positsioonil vähendab kromatiini avatust.



Joonis 11. Kromatiini avatusega seotud geeni PARP15 RNA sekveneerimise fragmentide kattuvus sõltuvalt kühmu juhtvariandi (rs2332287) genotüübist. Joonisel intronite ulatused vähendatud. Iga lisanduv tsütosiin juhtvariandi positsioonil suurendab ühe keskmise eksoni kaasamist.

Kromatiini avatuse piirkonna juhtvariant mõjutab r^2 analüüsil leitud seotud geeni PARP15 splaissimist (Joonis 11). Jooniselt on näha, et ühe keskmise eksoni kaasamine splaissimisel sõltub caQTL genotüübist. Iga lisanduv tsütosiin seotud kromatiini kühmu juhtvariandi positsioonil (rs2332287) suurendab ühe eksoni kaasamist lõplikku transkripti. Selline leid võib olla märgiks, et lisanduv tsütosiin caQTL positsioonil toob kaasa nõrgema splaiss-saidi valiku ja alternatiivse eksoni kaasamise.



Joonis 12. Ühe kromatiini avatuse kühmu ümbruses paiknevate lookuste nominaalsed p-väärtused, mis on leitud kromatiini avatuse ja splaissimise QTL analüüsil (caQTL juhtvariant rs232287, sQTL juhtvariant rs17208928).

Kromatiini avatuse QTL analüüsis olen leidnud 3. kromosoomil paikneva ligipääsetava kromatiini piirkonna jaoks vähima p-väärtusega variandi rs232287 (nominaalne p-väärtus = $6 * 10^{-11}$). Splaissimise QTL analüüsis leiti geeni PARP15 ühe eksoni kaasamist mõjutav juhtvariant rs17208928 (nominaalne p-väärtus = 10^{-90}). Lisaks leitud juhtvariantidele paikneb ümbruses mitmeid aheldatud variante (Joonis 12). Aheldatuse tõttu on variantidel ka

sarnane lineaarse mudeli p-väärtus ja see raskendab põhjusliku variandi kindlaks tegemist. Kromatiini avatuse ja splaissimise juhtvariantide aheldatuse määr on r^2=0,84. Eeldades, et igal geneetilisel variandil on väike tõenäosus olla põhjuslik variant, võib üsna suure tõenäosusega kujutatud näites kromatiini ja splaissimist mõjutada sama põhjuslik variant. Siiski pole võimalik selles lõplikult kindel olla. DNA avatust ja splaissimist võivad mõjutada erinevad põhjuslikud variandid, mis on lihtsalt tugevalt aheldatud. See on üldiselt geneetilise lähenemise piirang.

2.4.2. CTCFi ja HP1 mõju splaissimisele

Olen leidnud lookused, mis avaldavad mõju nii kromatiini avatusele kui varieeruvusele transkripti tasemel. Leidmaks tõendeid statistilise seose taga olevatest bioloogilistest regulatsioonimehhanismidest, uurin lähemalt transkriptsioonifaktorite CCCTC-seonduva faktori (CTCF) ja heterokromatiini valk 1 (HP1) seondumist. CTCF ja HP1 seost alternatiivse splaissimisega on varem näidatud erinevates rakuliinides (Yearim *et al*., 2015) (Agirre *et al*., 2015) (Shukla *et al*., 2011) (Ruiz-Velasco *et al*., 2017).

Statistiliselt olulise juhtvariandiga avatud kromatiini piirkonnad kattusid valdavas enamuses (98%) eelnevalt annoteeritud regulatoorsete aladega (annotatsioonid Ernst *et al*., 2011) ning 6,9% sisaldas CTCF seondumiskohta (ENCODE andmestiku järgi 10%). 3,3% huvipakkuvatest avatud kromatiini piirkondadest sisaldas HP1 seondumiskohta. Geneetiline variant, mis mõjutab kromatiini avatust ja splaissimist võib toimida läbi transkriptsioonifaktorite CTCF ja HP1 seondumisvõime muutmise. Selle uurimiseks teen kindlaks, kas avatud kromatiin, mis on seotud RNA taseme tunnusega, sisaldab suurema tõenäosusega transkriptsioonifaktori seondumiskohta. Esmalt jaotan avatud kromatiini piirkonnad kaheks selle põhjal, kas kromatiini piirkonna juhtvariant on seotud RNA taseme tunnusega või mitte. Seose avatud kromatiini kühmu ja RNA taseme tunnuse vahel olen kindlaks teinud r^2 analüüsil, mis võimaldab öelda, et tõenäoliselt mõjutab mõlema tunnuse varieeruvust sama põhjuslik variant.

Ootuspäraselt on kõigi huvipakkuvate avatud piirkondade hulgas transkriptsiooni või splaissimisega seotud piirkondade osakaal väiksem kui transkriptsioonifaktori seondumiskohta sisaldavate avatud piirkondade hulgas (Tabel 4). Kasutan Fisheri testi, et teha kindlaks, kas erinevus on statistiliselt oluline. Transkriptsiooni või splaissimisega seotud avatud kromatiini kühmude osakaal on suurem CTCF seondumiskohtade hulgas (r^2 > 0,8, Fisher p=0,001762939) (r^2 > 0,9, Fisher p=0,007996331) (Yurii Toma andmestik). ENCODE

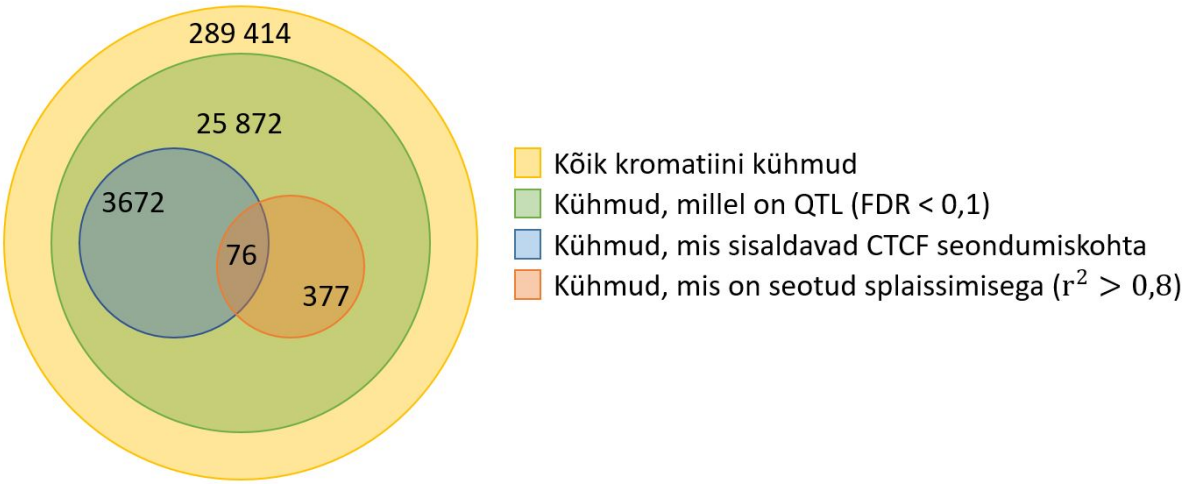
CTCF ja HP1 seondumiskohtade puhul pole erinevus statistiliselt oluline. Kuigi Toma andmestikus on erinevus statistiliselt oluline on vahe tegelikult üsna väike (8,7% vs 7,3%). Protsentuaalselt on ENCODE ja Toma andmed üsna sarnased ja võib öelda, et need kaks CTCF seondumiskohtade andmestikku on kooskõlas. Seega ei saa väita, et transkripti kasutusega seotud avatud kromatiinis oleks CTCF või HP1 seondumiskohad üleesindatus.

Tabel 4. Selliste kromatiini kühmude arv, millel leidub statistiliselt oluline QTL. Aheldatuse analüüsil on leitud kromatiini kühmud, mille juhtvariant mõjutab suure tõenäosusega ka vähemalt ühte RNA tunnust. Tabelis on CTCF või HP1 seondumiskohtade kattumine nende kühmudega.

	Seotud RNA QTLga ($r^2 > 0,8$)		Ei ole seotud RNA QTLga ($r^2 > 0,8$)	
	Sisaldab TF seondumis-kohta	Ei sisalda TF seondumis-kohta	Sisaldab TF seondumis-kohta	Ei sisalda TF seondumis-kohta
CTCF (Toma) seondumine	325 (8,7%)	1610 (7,3%)	3423	20514
CTCF (ENCODE) seondumine	208 (8,1%)	1727 (7,4%)	2366	21571
HP1 (ENCODE) seondumine	73 (8,5%)	1862 (7,4%)	781	23156
	Seotud RNA QTLga($r^2 > 0,9$)		Ei ole seotud RNA QTLga ($r^2 > 0,9$)	
	Sisaldab TF seondumis-kohta	Ei sisalda TF seondumis-kohta	Sisaldab TF seondumis-kohta	Ei sisalda TF seondumis-kohta
CTCF (Toma) seondumine	216 (5,8%)	1064 (4,8%)	3532	21060
CTCF (ENCODE) seondumine	139 (5,4%)	1141 (4,9%)	2435	22157
HP1 (ENCODE) seondumine	50 (5,9%)	1230 (4,9%)	804	23788

Järgmisena vaatlen avatud kromatiinis paiknevaid CTCF ja HP1 seondumiskohti ja nende mõju iga RNA tunnuse jaoks eraldi. Uurin, kas transkriptsioonifaktori seondumiskoht avatud kromatiinis väljendub RNA tasemel molekulaarses fenotüübis. CTCF või HP1 seondumiskohaga kattumine ei mõjutanud splaissimisega seotud kühmude osakaalu (Joonis

13). ENCODE andmestiku CTCF seondumiskohtade puhul ei esinenud ühtegi statistiliselt olulist erinevust (Lisa 1).



Joonis 13. Avatud kromatiini seos CTCF seondumise (Toma andmed) ja splaissimisega. Siin kujutatud jaotus on aluseks Fisheri täpsele testile. Testin, kas CTCF seondumisel on oluline roll splaissimise regulatsioonis.

Kromatiini avatuse kühmud, mis sisaldavad HP1 seondumiskohta on suurema tõenäosusega seotud geeniekpressiooniga (r^2 > 0.8, Fisher p=0,04559931) (r^2 > 0.9, Fisher p=0,03751518) (Lisa 2). Statistiliselt olulise erinevuse leidsin ka CTCF seondumise ja puQTLde puhul (r^2 > 0.8, Fisher p=0,027936823) (r^2 > 0.9, Fisher p=0,025947460) (Tabel 5). Kuid ka siin on protsentuaalne erinevus üsna väike (1,5% vs 1,1%).

Need tulemused võivad kinnitada, et avatud kromatiin määrab pigem geenieksprессiooni taseme ja promootori valiku kui splaissimise. Siiski ei ole avatud kromatiinis paikneval CTCF või HP1 seondumiskohal märkimisväärset mõju RNA tunnustele.

Tabel 5. Selliste kromatiini kühmude arv, millel leidub statistiliselt oluline QTL. Aheldatuse analüüsil on leitud kromatiini kühmud, mille juhtvariant mõjutab suure tõenäosusega ka RNA tunnuseid. Tabelis on Toma andmestiku CTCF seondumiskohtade kattumine nende ATAC kühmudega. Lahtrite värvid vastavad joonisele 13, kuid tabelis on esitatud tulemused mõlema r^2 taseme jaoks.

RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0,8$)		Ei ole seotud RNA QTLga ($r^2 > 0,8$)	
	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta
Ekspressioon	228 (6,1%)	1201 (5,4%)	3520	20923
Promootor	56 (1,5%)	245 (1,1%)	3692	21879
Splaissimine	76 (2,0%)	377 (1,7%)	3672	21747
3’ otsa kasutus	58 (1,5%)	283 (1,3%)	3690	21841
RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0,9$)		Ei ole seotud RNA QTLga ($r^2 > 0,9$)	
	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta
Ekspressioon	154 (4,1%)	802 (3,6%)	3594	21322
Promootor	35 (0,9%)	139 (0,6%)	3713	21985
Splaissimine	49 (1,3%)	223 (1,0%)	3699	21901
3’ otsa kasutus	33 (0,9%)	166 (0,8%)	3715	21958

Edasi vaatlen lookuseid, millel on statistiline seos CTCF seondumisega (Yurii Toma andmed) (Tabel 6). Kokku 49 698 CTCF seondumiskoha hulgas leidub 1112 seondumiskohta, millel on statistiline seos (FDR < 0,1) mõne lähedal asuva lookusega. Need CTCF seondumisele mõju avaldavad lookused on CTCF QTLd.

Üritan välja selgitada, kas geneetilised variandid, mis mõjutavad CTCF seondumist on olulised ka splaissimise regulatsioonis. Selleks kasutan r^2 analüüsi, et leida kromatiini kühmud, splaissimissündmused ja CTCF seondumiskohad, mille varieeruvust mõjutab suure tõenäosusega ühine põhjuslik variant. Leidsin 474 kromatiini avatuse kühmu, millel on seos mõne CTCF seondumiskohaga ja 453 kühmu, millel on seos mõne geeni splaissimisega (r^2 >

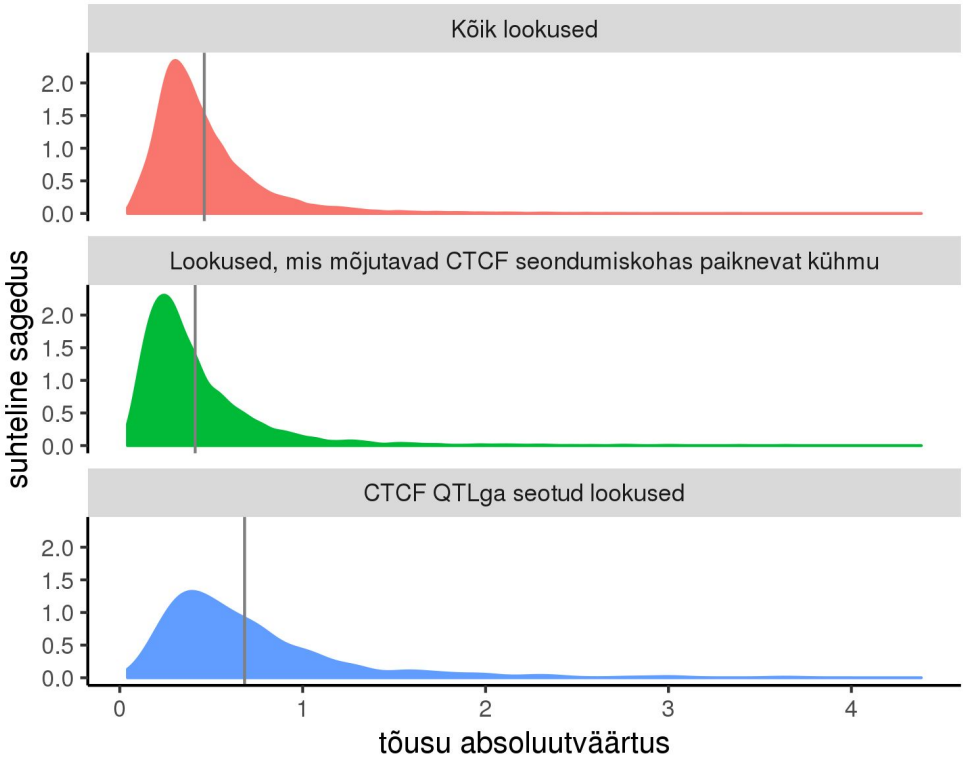
0,8). Sama r^2 taseme juures leidsin 40 kromatiini avatuse kühmu, mis on seotud nii splaissimise kui ka CTCF seondumisega.

Lookused, mis on seotud kromatiini avatuse ja CTCF seondumisega mõjutavad suurema tõenäosusega splaissimist ($r^2 > 0.8$, Fisher $p=9,8*10^{-8}$) kui kõik kromatiini avatusega seotud lookused. Sarnase tulemuse leidsin ka geenieksрessiooni, promootori kasutuse ja 3' otsa valikuga seotud lookuste jaoks mõlema r^2 taseme jaoks. Siiski ei tuvastanud, et CTCF QTL mõjutaks suurema tõenäosusega splaissimist kui mõnda teist RNA taseme tunnust. Igat tüüpi RNA taseme tunnusega seotud kromatiini kühmude hulgas leidus sarnane (erinevus polnud statistiliselt oluline) osakaal ülekatet CTCF lookustega.

Tabel 6. Selliste kromatiini kühmude arv, millel leidub statistiliselt oluline QTL. Aheldatuse analüüsil on leitud kromatiini kühmud, mille juhtvariant mõjutab suure tõenäosusega ka CTCF seondumist (CTCF QTLde põhjal) ja RNA tunnuseid.

RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0.8$)		Ei ole seotud RNA QTLga ($r^2 > 0.8$)	
	Seotud CTCF QTLga	Ei ole seotud CTCF QTLga	Seotud CTCF QTLga	Ei ole seotud CTCF QTLga
Ekspressioon	108 (22,8%)	1321 (5,2%)	366	24077
Promootor	21 (4,4%)	280 (1,1%)	453	25118
Splaissimine	27 (5,7%)	426 (1,7%)	447	24972
3' otsa kasutus	31 (6,5%)	310 (1,2%)	443	25088
üle kõigi tüüpide	127 (26,8%)	1808 (7,1%)	347	23590
RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0.9$)		Ei ole seotud RNA QTLga ($r^2 > 0.9$)	
	Seotud CTCF QTLga	Ei ole seotud CTCF QTLga	Seotud CTCF QTLga	Ei ole seotud CTCF QTLga
Ekspressioon	49 (16,0%)	907 (3,6%)	257	24659
Promootor	7 (2,3%)	167 (0,7%)	299	25399
Splaissimine	14 (4,6%)	258 (1,0%)	292	25308
3' otsa kasutus	7 (2,3%)	192 (0,8%)	299	25374
üle kõigi tüüpide	62 (20,2%)	1218 (4,8%)	244	24348

Olen jõudnud mõneti vastuoluliste tulemusteni. Avatud kromatiinis paiknev CTCF seondumiskoht ei ole määrav splaissimise regulatsioonis. Samas on CTCF QTLd suurema tõenäosusega seotud splaissimise regulatsiooniga. CTCF QTLd leiti väikese valimi põhjal (n = 49) (Toma, 2018). Väikese valimi suuruse tõttu on pärast mitmese testimise korrektuure statistiliselt olulised ainult suure mõjuga lookused. Lookuse efektisuurust fenotüübi tunnusele hindab regressioonimudeli sirge tõus. Suurema tõusu absoluutväärtuse korral on lookusel tunnuse varieeruvusele suurem efekt. Kromatiini avatuse kühmude juhtvariandid, mis on aheldatud CTCF QTLga on keskmisest suurema efektisuurusega (Joonis 14).



Joonis 14. QTL analüüsi regressioonisirge tõus hindab lookuse efektisuurust fenotüübi tunnusele. Iga jaotuse graafiku alune pindala 1. Keskväärtus märgitud halli joonega.

Väikese valimi suuruse juures leitud CTCF QTLd võivad tähistada lookuseid, millel on üldiselt suur efekt molekulaarsele fenotüübile. Suurema efektisuurusega kaasneb suurem statistiline võimsus, et tuvastada ka nende efekt geeniekpressiooni või splaissimise regulatsioonis. Leidsin 3748 CTCF seondumiskohta sisaldavat avatud kromatiini piirkonda, millel leidub statistiliselt oluline caQTL. Samas ainult 474 juhtumit, kus CTCF seondumist ja kromatiini avatust mõjutab suure tõenäosusega sama geneetiline variant. Kuna transkriptsioonifaktorite kinnituskohaks on ainult avatud kromatiin siis peaks kõik CTCF

seondumiskohta sisaldavate avatud piirkondade juhtvariandid mõjutama mingil määral ka CTCF seondumist.

Kõigi seoste tuvastamiseks puudub piisav statistiline võimsus. Seetõttu on fenotüübi tunnuseid mõjutavate geneetiliste variantide otsimisel palju valenegatiivseid tulemusi. See võib olla põhjuseks, miks CTCF seondumiskohta sisaldav avatud kromatiin ei ole märkimisväärselt suurema tõenäosusega seotud splaissimise kontrolliga (2% vs 1,7%). Samal ajal kui caQTLd, mis on ka CTCF QTLd on suurema tõenäosusega seotud splaissimisega kui kõik caQTLd (4,6% vs 1%).

Faktorite seondumisest splaissimiseni on palju samme. Splaissimise regulatsiooni mõjutavad veel muud tegurid, mida käesolevas töös pole käsitletud. ChIP-seq analüüsis leiti, et CTCF mõju splaissimisele väljendub juhul kui CTCF seondumiskoht paikneb promootori ja eksoni vahel (Ruiz-Velasco *et al*., 2017). Mitmes analüüsis leiti, et HP1 ja CTCF mõju splaissimisele võib sõltuda DNA metüleeritusest (Yearim *et al*., 2015) (Hashimoto *et al*., 2017). Käesolevas töös pole kasutatud metüleerituse andmeid. Seega pole võimalik kinnitada metüleerituse rolli eksonite kaasamisel või kõrvale jätmisel. RNA polümeraas II ChIP-seq signaali põhjal leiti, millise mehhanismi läbi faktorid splaiss-saitide valikut mõjutavad (Agirre *et al*., 2015). Käesolevas töös on transkriptsioonifaktorite seondumiskohtade põhjal võimalik uurida vaid statistilisi seoseid.

Splaissimise regulatsioon toimub läbi keerulise RNA ja valkude võrgustiku. Selles regulatsioonis osalevad DNAle seonduvad valgud CTCF ja HP1. Regulatsioonimehhanismid on keerulised ning käesolevas töös tehtud analüüsi põhjal ei olnud võimalik kindlaks teha põhilisi mehhanisme, mille läbi avatud kromatiin mõjutab splaissimist. Selgemaid tulemusi võiks saavutada suurema valimi korral, sest see tooks kaasa suurema statistilise võimsuse seoste avastamiseks. Edasistes analüüsides tuleks arvestada ka tegureid nagu metüleeritus ja RNA polümeraasi kinnitumine või faktori seondumiskoha täpne paiknemine reguleeritava geeni suhtes.

Kokkuvõte

ATAC sekveneerimisel mõõdetud kromatiini ligipääsetavuse andmed olid kooskõlas varasemate DNase I sekveneerimise tulemustega. Kuna DNase I sekveneerimiseks vajalik laboriprotokoll on tunduvalt aeganõudvam siis on edasistes analüüsides sobilik kasutada just ATAC sekveneerimist. Käesolevas töös õnnestus leida geneetilisi variante, mis suure

tõenäosusega mõjutavad samaaegselt kromatiini avatust ja alternatiivset splaissimist Epstein-Barri viirusega nakatatud B-rakkudes. Statistilist seost kinnitavad sekveneerimislugemite kattuvuse joonised, millel võib vaadelda kromatiini avatuse varieeruvuse ja alternatiivse splaissimise lineaarset seost ühe lookuse genotüübiga. Siiski ei tundu avatud kromatiin olevat põhiline splaissimisega seotud lookuste määraja. Kromatiini avatusega seotud lookused on määravad pigem geenieksрessiooni ja promootori valiku kui splaissimise regulatsioonis.

Töös ei õnnestunud välja selgitada, milline on põhiline mehhanism, kuidas üks geneetiline variant mõjutab nii kromatiini avatust kui ka splaissimist. CTCF või HP1 seondumine avatud kromatiinile ei tundu olevat põhiliseks mehhanismiks. CTCF või HP1 seondumiskohta sisaldavate avatud kromatiini piirkondade hulgas pole märkimisväärselt suurem osakaal splaissimisega seotud piirkondi kui kõigi avatud piirkondade hulgas. Lookused, mis on caQTLd ja CTCF QTLd on suurema tõenäosusega seotud splaissimisega kui kõik caQTLd.

Selle statistilise seose taga võib olla CTCF roll splaissimise regulaatorina või leitud CTCF QTLde suurem efektisuurus kogu molekulaarsele fenotüübile, mis tuleneb analüüsi meetodist. Lookused, mis on caQTLd ja CTCF QTLd on suurema tõenäosusega seotud ka geenieksрessiooni, promootori valiku ja 3’ otsa valikuga. Seega ei õnnestunud tuvastada CTCF erilist rolli just splaissimise regulatsioonis. Edasistes kromatiini avatuse ja splaissimise analüüsides tuleks arvestada täiendavaid tegureid, millel on varem näidatud rolli splaissimise regulatsioonis.

Summary

Discovering genetic variants that affect both chromatin accessibility and RNA splicing

Evelin Aasna

Summary

Splicing has an important role in contributing to phenotypic variation. The number of different proteins needed to sustain an organism greatly exceeds the number of protein coding genes. Variability of protein products is increased through alternative splicing of RNA transcripts. RNA splicing occurs in the majority of human genes and mutations affecting splicing have been implicated in many genetic diseases. How genetic differences contribute to phenotypic traits is a central question of genome research.

It has been shown that gene expression and splicing are controlled by mostly independent sets of genetic variants. Previous analyses of lymphoblastoid cell lines (LCLs) has identified an overlap between genetic variants that affect chromatin accessibility and RNA splicing. Using ATAC and RNA sequencing data, I detected genetic variants that likely affect both chromatin accessibility and RNA splicing in LCL cells. However, genetic variants that affect chromatin accessibility were not the main drivers of splicing regulation. They were more likely to have an effect on promoter selection and gene expression levels.

Splicing is genetically regulated by an intricate network of proteins and RNA molecules. A genetic variant can affect both chromatin accessibility and RNA splicing through affecting the binding of factors that regulate splicing. Previous studies have shown that accessible chromatin and RNA splicing might be linked through CCCTC binding factor (CTCF) and heterochromatin protein 1 (HP1). I used CTCF and HP1 binding sites to see if these factors have a central role in mediating splicing regulation through binding to DNA.

Finding a CTCF or HP1 binding site in accessible chromatin was not a good indicator that the lead variant affecting chromatin accessibility would also be linked to splicing of nearby genes. Genetic variants that affected both chromatin accessibility and CTCF binding were more likely to also affect gene expression or splicing of nearby genes. This statistically significant result might be due to the CTCF associated genetic variants having an overall stronger effect on molecular phenotypes. I did not detect a splicing specific effect of CTCF associated variants.

Given previous work, it is likely that binding of CTCF and HP1 to accessible chromatin does have a role in splicing regulation. However, I did not detect a substantial role of CTCF and HP1 binding in regulating splicing. This analysis could not take into account the full complexity of splicing regulation. Further analyses considering additional factors such as methylation and RNA polymerase binding might be more successful in uncovering the mechanisms by which a genetic variant can affect both chromatin accessibility and RNA splicing.

Kirjanduse loetelu

Agirre, E., Bellora, N., Alló, M., Pagès, A., Bertucci, P., Kornblihtt, A.R., and Eyras, E. (2015). A chromatin code for alternative splicing involving a putative association between CTCF and HP1α proteins. BMC Biol. 13, 31.

Alasoo, K., and Fishman, D. Predicting the impact of non-coding genetic variants on transcription factor binding with machine learning. 47.

Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D.F., Paul, D.S., and Gaffney, D.J. (2018). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. BioRxiv 319806.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16: 197–212.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature 465: 53–59.

Breathnach, R., and Chambon, P. (1981). Organization and Expression of Eucaryotic Split Genes Coding for Proteins. Annu. Rev. Biochem. 50: 349–383.

Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 109: 21.29.1-21.29.9.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213–1218.

Davis, C.A., Hitz, B.C., Sloan, C.A., ... Cherry, J.M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46: D794–D801.

Degner, J.F., Pai, A.A., Pique-Regi, R., ... Pritchard, J.K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390–394.

Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat. Commun. 8, 15452.

Ding, Z., Ni, Y., Timmer, S.W., ... Birney, E. (2014). Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. PLOS Genet. 10, e1004798.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., ... Berntein, B.E. (2011). Mapping and analysis of

chromatin state dynamics in nine human cell types. Nature 473: 43–49.

Felsenfeld, G., Boyes, J., Chung, J., Clark, D., and Studitsky, V. (1996). Chromatin structure and gene expression. Proc. Natl. Acad. Sci. U. S. A. 93: 9384–9388.

Fort, A., Panousis, N.I., Garieri, M., Antonarakis, S.E., Lappalainen, T., Dermitzakis, E.T., and Delaneau, O. (2017). MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. Bioinforma. Oxf. Engl. 33: 1895–1897.

Gaut, B.S., and Long, A.D. (2003). The Lowdown on Linkage Disequilibrium. Plant Cell 15: 1502–1506.

Gross, D.S., and Garrard, W.T. (1988). Nuclease Hypersensitive Sites in Chromatin. Annu. Rev. Biochem. 57: 159–197.

Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., ... Dermitzakis, E.T. (2015). Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet. 11.

Hansen, K.D., Irizarry, R.A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostat. Oxf. Engl. 13: 204–216.

Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G., and Cheng, X. (2017). Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol. Cell 66: 711-720.e3.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15, 182.

Johannsen, W. The Genotype Conception of Heredity. Am. Nat. 31.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science 184: 868–871.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522.

Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. 48: 206–213.

Lander, E.S., and Schork, N.J. Genetic Dissection of Complex Traits. 12.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ... Dermitzakis, E.T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,

and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25: 2078–2079.

Li, Y.I., Geijn, B. van de, Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science 352: 600–604.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. 30: 923–930.

Linker, S.M., Urban, L., Clark, S.J., ... Bonder, M.J. (2019). Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. Genome Biol. 20, 30.

Martin, A., and Orgogozo, V. (2013). The Loci of Repeated Evolution: A Catalog of Genetic Hotspots of Phenotypic Variation. Evolution 67: 1235–1250.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. 6: 386–398.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457–463.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40: 1413–1415.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10: 669–680.

Pritchard, J.K., and Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data. Am. J. Hum. Genet. 69: 1–14.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Raha, D., Hong, M., and Snyder, M. (2010). ChIP-Seq: A Method for Global Identification of Regulatory Elements in the Genome. Curr. Protoc. Mol. Biol. 91: 21.19.1-21.19.14.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. 7, 862.

Ruiz-Velasco, M., Kumar, M., Lai, M.C., Bhat, P., Solis-Pinson, A.B., Reyes, A., Kleinsorg, S., Noh, K.-M., Gibson, T.J., and Zaugg, J.B. (2017). CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. Cell Syst. 5:

628-637.e6.

Schor, I.E., Gómez Acuña, L.I., and Kornblihtt, A.R. (2013). Coupling between transcription and alternative splicing. Cancer Treat. Res. 158: 1–24.

Sharp, P.A. (1988). RNA Splicing and Genes. JAMA 260: 3035–3041.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479: 74–79.

Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. 2010, pdb.prot5384.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526: 68–74.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Thurman, R.E., Rynes, E., Humbert, R., ... Stamatoyannopoulos, J.A. (2012). The accessible chromatin landscape of the human genome. Nature 489: 75–82.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics Chromatin 7.

Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. 8: 749–761.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA 14: 802–813.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57–63.

Welter, D., MacArthur, J., Morales, J., ... Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42: D1001–D1006.

Yearim, A., Gelfman, S., Shayevitch, R., ... Ast, G. (2015). HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Rep. 10: 1122–1134.

Zhang, Y., Liu, T., Meyer, C.A., ... Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

Kasutatud veebiaadressid

http://www.internationalgenome.org/data 2018

https://www.encodeproject.org/ 2018

http://broadinstitute.github.io/picard/ 2018

https://www.encodeproject.org/software/bedgraphtobigwig/ 2018

https://github.com/kauralasoo/Blood_ATAC/blob/master/Snakefile 2018

http://bioconductor.org/packages/release/bioc/html/wiggleplotr.html 2019

https://www.ncbi.nlm.nih.gov/assembly?term=GRCh38&cmd=DetailsSearch 2018

Lisad

Lisa 1. Selliste kromatiini kühmude arv, millel leidub statistiliselt oluline QTL. Aheldatuse analüüsil on leitud kromatiini kühmud, mille juhtvariant mõjutab suure tõenäosusega ka RNA tunnuseid. Tabelis on ENCODE andmestiku CTCF seondumiskohtade kattumine nende ATAC kühmudega.

RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0,8$)		Ei ole seotud RNA QTLga ($r^2 > 0,8$)	
	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta
Ekspressioon	227 (6,1%)	1170 (5,3%)	3521	20954
Promootor	56 (1,5%)	245 (1,1%)	3692	21879
Splaissimine	76 (2,0%)	377 (1,7%)	3672	21747
3’ otsa kasutus	58 (1,5%)	283 (1,3%)	3690	21841
RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0,9$)		Ei ole seotud RNA QTLga ($r^2 > 0,9$)	
	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta	Sisaldab CTCF seondumis-kohta	Ei sisalda CTCF seondumis-kohta
Ekspressioon	100 (3,9%)	856 (3,7%)	2474	22442
Promootor	23 (0,9%)	151 (0,6%)	2551	23147
Splaissimine	29 (1,1%)	243 (1,0%)	2545	23055
3’ otsa kasutus	21 (0,8%)	178 (0,8%)	2553	23120

Lisa 2. Selliste kromatiini kühmude arv, millel leidub statistiliselt oluline QTL. Aheldatuse analüüsil on leitud kromatiini kühmud, mille juhtvariant mõjutab suure tõenäosusega ka RNA tunnuseid. Tabelis on ENCODE andmestiku HP1 seondumiskohtade kattumine nende ATAC kühmudega.

RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0.8$)		Ei ole seotud RNA QTLga ($r^2 > 0.8$)	
	Sisaldab HP1 seondumis-kohta	Ei sisalda HP1 seondumis-kohta	Sisaldab HP1 seondumis-kohta	Ei sisalda HP1 seondumis-kohta
Ekspressioon	59 (6,9%)	1370 (5,5%)	795	23648
Promootor	11 (1,3%)	290 (1,2%)	843	24728
Splaissimine	13 (1,5%)	440 (1,8%)	841	24578
3’ otsa kasutus	6 (0,7%)	335 (1,3%)	848	24683
RNA QTL tüüp	Seotud RNA QTLga ($r^2 > 0.9$)		Ei ole seotud RNA QTLga ($r^2 > 0.9$)	
	Sisaldab HP1 seondumis-kohta	Ei sisalda HP1 seondumis-kohta	Sisaldab HP1 seondumis-kohta	Ei sisalda HP1 seondumis-kohta
Ekspressioon	42 (4,9%)	914 (3,7%)	812	24104
Promootor	5 (0,6%)	169 (0,7%)	849	24849
Splaissimine	9 (1,1%)	263 (1,1%)	845	24755
3’ otsa kasutus	3 (0,4%)	196 (0,8%)	851	24822

Lihtlitsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Evelin Aasna,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Samaaegselt kromatiini avatust ja splaissimist mõjutavate geneetiliste variantide leidmine", mille juhendaja on PhD Kaur Alasoo, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Evelin Aasna
27.05.2019